Research Topics

Computational sociolinguistics

Edited by
Jack Grieve, Dirk Hovy, David Jurgens, Tyler S. Kendall,
Dong Nguyen, James N. Stanford and Meghan Sumner

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-1760-4
DOI 10.3389/978-2-8325-1760-4

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational sociolinguistics

Topic editors

Jack Grieve — University of Birmingham, United Kingdom
Dirk Hovy — Bocconi University, Italy
David Jurgens — University of Michigan, United States
Tyler S. Kendall — University of Oregon, United States
Dong Nguyen — The Alan Turing Institute, United Kingdom
James N. Stanford — Dartmouth College, United States
Meghan Sumner — Stanford University, United States

Citation

Grieve, J., Hovy, D., Jurgens, D., Kendall, T. S., Nguyen, D., Stanford, J. N., Sumner, M., eds. (2023). *Computational sociolinguistics*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-1760-4

Table of contents

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 20 June 2019
doi: 10.3389/frai.2019.00010

Sources of Microtemporal Clustering in Sociolinguistic Sequences

*Meredith Tamminga**

Department of Linguistics, University of Pennsylvania, Philadelphia, PA, United States

Persistence is the tendency of speakers to repeat the choice of sociolinguistic variant they have recently made in conversational speech. A longstanding debate is whether this tendency toward repetitiveness reflects the direct influence of one outcome on the next instance of the variable, which I call sequential dependence, or the shared influence of shifting contextual factors on proximal instances of the variable, which I call baseline deflection. I propose that these distinct types of clustering make different predictions for sequences of variable observations that are longer than the typical prime-target pairs of typical corpus persistence studies. In corpus ING data from conversational speech, I show that there are two effects to be accounted for: an effect of how many times the /ing/ variant occurs in the 2, 3, or 4-token sequence prior to the target (regardless of order), and an effect of whether the immediately prior (1-back) token was /ing/. I then build a series of simulations involving Bernoulli trials at sequences of different probabilities that incorporate either a sequential dependence mechanism, a baseline deflection mechanism, or both. I argue that the model incorporating both baseline deflection and sequential dependence is best able to produce simulated data that shares the relevant properties of the corpus data, which is an encouraging outcome because we have independent reasons to expect both baseline deflection and sequential dependence to exist. I conclude that this exploratory analysis of longer sociolinguistic sequences reflects a promising direction for future research on the mechanisms involved in the production of sociolinguistic variation.

Keywords: sociolinguistics, persistence, priming, style-shifting, simulation, corpus

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Ryan L. Boyd,
University of Texas at Austin,
United States
Anastasia Smirnova,
San Francisco State University,
United States

Correspondence:
Meredith Tamminga
tamminga@ling.upenn.edu

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *19 March 2019*
Accepted: *29 May 2019*
Published: *20 June 2019*

Citation:
Tamminga M (2019) Sources of
Microtemporal Clustering in
Sociolinguistic Sequences.
Front. Artif. Intell. 2:10.
doi: 10.3389/frai.2019.00010

1. INTRODUCTION

Quantitative sociolinguists have long known that in conversational speech, speakers tend to repeat the choice of the sociolinguistic variant they have recently made. Following Szmrecsanyi (2006), I call this phenomenon *persistence*[1]. Persistence has been observed for a wide range of variables across multiple languages, including pronominal alternations in Quebec French (Sankoff and Laberge, 1978), the passive alternation in English (Weiner and Labov, 1983; Estival, 1985), /s/-deletion and /n/-deletion in Puerto Rican Spanish (Poplack, 1980, 1984), verbal /s/ omission in some varieties of English (Poplack and Tagliamonte, 1989), /s/-deletion in Brazilian Portuguese (Scherre and Naro, 1991, 1992; Scherre, 2001), the English dative alternation (Gries, 2005), particle placement in English (Gries, 2005; Szmrecsanyi, 2006), English coronal stop deletion (Tamminga, 2016), and more. The evidence is abundant that a speaker's choice of variant for a variable at any given moment is partly predictable from their most recent variant choice for the same variable.

[1] It has also sometimes been called perseverance, perseveration, serialism, parallelism, and most colorfully, the "birds of a feather" effect.

How to explain this phenomenon, though, is more controversial. Broadly speaking, there are two classes of explanation. Tamminga et al. differentiate between *sequential dependence*, which is when "the outcome of a sociolinguistic alternation in one moment directly influences the likelihood of a matching outcome some moments later" (Tamminga et al., 2016, p. 33), and *baseline deflection*, which is when "two closely-proximal instances of a sociolinguistic variable are more likely to occur under similar social-contextual circumstances than two instances that are further apart, and thus are more likely to have matching outcomes" (Tamminga et al., 2016, p. 34). Both of these could in principle produce the kind of microtemporal clustering that has been called persistence. Research on persistence sometimes assumes sequential dependence and attributes the dependence to priming, in the psycholinguistic sense of facilitated access to a recently encountered linguistic form [2]. But it has also been repeatedly observed that stylistic forces might produce apparently similar repetitiveness. To trace an example in the literature, Weiner and Labov (1983) find that speakers are more likely to choose a passive construction instead of an active one when they have already recently used a passive. Weiner and Labov attribute this to both a "mechanical tendency to preserve parallel structure" (suggesting sequential dependence) and "a stylistic factor operating" (suggesting baseline deflection) (Weiner and Labov, 1983, p. 56). In a subsequent study building on Weiner and Labov's results, Estival concludes that "the effect we have been studying [is] a syntactic priming effect" (Estival, 1985, p. 21). In other words, she asserts that persistence in the passive involves sequential dependence in the form of structural priming. On the other hand, Branigan et al. raise the possibility of baseline deflection when they point out that Weiner and Labov's result "might just reflect shifts in the register used during the interviews which they studied" (Branigan et al., 1995, p. 492). Distinguishing between these possibilities is not straightforward.

In this paper I propose that we can make some progress in disentangling sequential dependence and baseline deflection by looking at sequences of multiple observations of the variable prior to a target instance of that variable, instead of just the immediately prior observation. These sequences reflect a string of prior instances on which the speaker had to choose between two[3] variants of the same sociolinguistic variable as in the target, each of which may be separated by some distance from the target and from other prior observations. For a variable with two possible variants A and B, the usual approach to persistence is to ask whether the probability of choosing B at target T is different based on whether the prior token was A or B: does the outcome in what I will call the A-T and B-T conditions differ?[4] If we extend our view back to the *two* choices the speaker made before the target, it will give us four conditions: A-A-T, B-A-T, A-B-T, and B-B-T [5]. I call this a 2-prior sequence, and say that the B-A-T sequence has a 1-back variant of A and a 2-back variant of B (that is, I use "2-prior" to refer to the total depth of the sequence before the target, and "2-back" to refer to a single observation in a particular position within the sequence). We can then ask how the probability of getting B at the target T differs in those four conditions. For instance, we might hypothesize that the observed rate of B in the target will be higher in the A-B-T condition than the B-A-T condition because in A-B-T, the prior instance of B occupies a slot closer in the sequence to the target.

In section 2.3, I conduct this type of quantitative analysis on 2-prior, 3-prior, and 4-prior sequences for the variable ING[6] in conversational speech. ING is the alternation between the velar and alveolar nasal after unstressed /ɪ/, as in *working* vs. *workin'*. Previous work has attributed ING persistence to priming (Abramowicz, 2007; Tamminga, 2014, 2016), but this variable has also been shown to exhibit style-shifting within data very comparable to that used here (Labov, 2001), making ING a suitable test case for this analysis. Both in section 2.3 and in further statistical analyses of the corpus data in section 2.4, I will demonstrate that the probability of the /ing/[7] variant is influenced by how many instances of /ing/ occur in the N-prior sequence, as well as by which variant occurs in the 1-back position. There is not, however, evidence that the probability of /ing/ in the target additionally depends on the ordering of the variants at a depth greater than 1-back.

After showing how N-prior sequences influence ING outcomes in the corpus data, I turn in section 3 to a series of simulations to explore what kind of process may have produced the patterns observed in speech. I create a series of simulations based on Bernoulli processes—in essence, modeling sociolinguistic variation as the flipping of weighted coins. The simulations can be set up to have different sources of microtemporal clustering built in, or to exclude such sources. One version of the simulation has sequential dependence built in, while others involve various simple versions of baseline deflection. With each simulation, I generate a dataset that can be analyzed using the same approach as I took with the corpus data, allowing for an intuitive comparison of the outcomes. While every simulation with any source of microtemporal clustering built in produces a difference of some magnitude based on the 1-prior sequence (that is, the analog to the usual persistence effect), the predicted probability as a function of the 3-prior sequence can differ more substantially between models containing baseline deflection and ones containing sequential dependence.

The possibilities for this type of simulation are enormous, and pursuing an exhaustive search of what it might produce is beyond the scope of such preliminary work as this paper. I will, however,

[2] Priming itself might arise from a variety of mechanisms, such as spreading activation or error-driven implicit learning, any of which would fall under the umbrella of sequential dependence.

[3] Or more, although I will not consider variables with more than two variants here.

[4] I explicitly include the T in the sequence name to make the directionality clear.

[5] Note that in these cases, the hyphens elide an unknown amount of speech between observations of the variable; in section 4 I will briefly address the question of the distance between observations, but I will mostly leave modeling of decay in multi-token sequences for later work.

[6] Following one variationist convention, the all-capitalized representation ING represents the variable itself, the choice between two outcome variants.

[7] I will use orthographic representations inside slashes for the variants: /ing/ for /ɪŋ/ and /in/ for /ɪn/ to achieve consistency with my sequence notation. For N-prior sequences, I put the entire sequence between a single pair of slashes. In graphs, I omit the slashes as unnecessary visual clutter. Although unconventional, I believe this is the most visually distinctive set of options, and therefore is to be preferred as a way of making the complex discussion slightly easier to follow.

suggest that each of the two mechanisms of microtemporal clustering maps more cleanly and consistently to one of the two central effects in the corpus data: baseline deflection can produce the effect of how many times /ing/ occurred in the prior sequence and sequential dependence straightforwardly gives rise to the effect of the immediately-prior token. The pattern seen in the corpus ING data, then, can be produced most effectively by a simulation in which I include both sequential dependence and baseline deflection mechanisms. I argue that this is a welcome result because there are independent reasons to believe in linguistic behavioral phenomena (as I discuss in the following subsection) that should give rise to both of these types of clustering. Finding out that their combination is necessary to produce observed microtemporal patterns in corpus data suggests that future work on persistence might move beyond either/or questions about the source of persistence.

1.1. A Terminological Note

The sociolinguistics and corpus linguistics literatures have often used the term "priming" for persistence. Objections to this designation have usually been framed in terms of "style-shifting" or "register changes." I will avoid using these terms throughout this paper even though the discussion would surely read more intuitively if I contrasted "priming" (sequential dependence) models with "style-shifting" (baseline deflection) models. However, I will maintain that the content- and context-blind quantitative modeling I will explore in this paper does not and cannot distinguish between different real-world interpretations of the microtemporal structures I am exploring. It is tempting to suggest that sequential dependence should be interpreted as the psychological effect of priming—which would itself still leave many questions about the priming mechanism unanswered. However, stylistic and discourse-structural considerations could also give rise to an effect of true sequential dependence. For instance, even if a choice of a particular word order alternant was made purely stochastically, unrelated to contextual preferences, a speaker might wish to continue with the same choice on later utterances in order to maintain the parallelism of the discourse. Similarly, speakers might tend toward repetitiveness itself as a stylistic choice rather than making a series of independent choices that happen to all be occurring under the influence of the same external situation. The same ambiguity is present when it comes to baseline deflection. It may seem most natural to understand shifts in a speaker's target variant rate as being the result of style-shifting, but it is also quite possible to think of psychological factors that could have a similar effect in jointly shaping sequences of target outputs. For instance, a speaker might be operating under a greater memory or attentional burden at some stretches of speech than others, which in turn might influence self-monitoring behavior. The quantitative approach taken here does not distinguish these possibilities; it only distinguishes between the quantitative properties of baseline deflection and sequential dependence. The evidence for how these distinct sources of microtemporal clustering should be interpreted will have to come from other directions. Most importantly, the evidence on this question of interpretation will need to come from conversational corpus data analysis that attends to speaker identity and behavior in particular sociointeractional contexts; such work might conceivably be supplemented by focused, socially sensitive experimental investigations.

2. PRIOR SEQUENCES OF THE ING VARIABLE

In previous work, I have shown that there is a relationship between a token of ING and the most recent token of ING from the same speaker (Tamminga, 2014, 2016), specifically that the speaker is likely to repeat their immediately prior variant choice. This is consistent with earlier work from Abramowicz (2007), as well as with the corpus persistence literature more generally. Here I use the same underlying dataset as in my previous work to extend my consideration of ING persistence to 2-prior, 3-prior, and 4-prior sequences.

2.1. Data

The conversational speech data come from the Philadelphia Neighborhood Corpus (PNC, Labov and Rosenfelder, 2011). The PNC contains sociolinguistic interviews recorded in Philadelphia between 1972 and 2012. The recordings have been orthographically transcribed, then automatically forced-aligned at the word and phone level using the FAVE-align component of the FAVE suite (Rosenfelder et al., 2011). The master ING dataset used here, which comes from a 118-speaker subset of the PNC, is the same as that described in Tamminga (2014, 2016); more detail on the speaker demographics can be found there. To create that dataset, I coded all of the ING observations in the sample auditorily using a Praat script to facilitate exhaustive searching of the corpus' FAVE-aligned TextGrids [8]. The data are coded with 0 representing /in/ and 1 representing /ing/, so values closer to 1 indicate a higher probability of the /ing/ variant being chosen. The data used for analysis in the current paper is a subset of this master ING dataset; details of how and why this particular subset was chosen are given in section 2.2 below. The primary predictor of interest in this study is the makeup of the N-prior sequence. Each ING token was coded for the values of the four prior ING observations from the same speaker, modulo the exclusions described in section 2.2. The multivariate analyses described in section 2.4 also include the following control predictors:

- Whole word frequency: the Lg10CD measure from SUBTLEX (Brysbaert and New, 2009)
- Speech rate: the number of vowels per second in a 7-word window centered on the target word, which is automatically collected by the Praat script originally used to code the data
- Preceding coronal: in this dataset ING shows progressive dissimilation
- Following pause: in this dataset /ing/ is more frequent before a pause
- Speaker gender: male or female, since ING is a classic stable variable, with women on average using more /ing/ than men.

[8] Thanks to Joe Fruehwald for sharing his handCoder.Praat script.

2.2. Revisiting the Envelope of Variation

In quantitative sociolinguistics, deciding what to count and how to count it is a crucial process, sometimes called defining the envelope of variation. I give special attention to these decisions here because, as I point out in Tamminga (2014), the study of persistence raises new issues for the envelope of variation. Two of these issues are relevant here: the role of the interlocutor and the definition of the variable itself.

Regarding the role of the interlocutor, in Tamminga (2014, 2016), I omit prime–target pairs that were interrupted by an instance of the variable from an interlocutor. The reason for this decision is that we do not currently know how phenomena like accommodation and interspeaker priming interact with intraspeaker persistence, so we should neither assume that an ING token from an interlocutor is the same as a token from the target speaker and can be included, nor assume that it is irrelevant and can be ignored. Here I extend that decision to the consideration of sequences, making interruption-based exclusions for the length of the N-prior sequence at hand. **Figure 1** illustrates that if there had been no interruption, the target at t_4 would have had a 3-prior sequence of /ing-ing-in-T/, while the target at t_3 would also have had a 2-prior sequence of /ing-ing-T/ and could have been included in a 2-prior analysis. But because there is an interruption between the 2-back and 3-back positions relative to the target at t_4, t_4 ends up with no 3-prior sequence, but does still have a valid 2-prior sequence of /ing-in-T/. With this practice, the number of targets that can be included is reduced at each greater depth of prior token sequence.

The second issue is that of the definition of the dependent variable itself. So far I have defined ING as the alternation between the velar and alveolar nasal after unstressed /ɪ/, but complications arise because this alternation occurs in a range of grammatical contexts. Often the ING variable is defined as including progressive verbs and gerunds formed with the *-ing* suffix, such as *working*, monomorphemes like *ceiling*, and the words *something* and *nothing*. However, there has long been uncertainty about whether or not the surface variability in these contexts is the output of a single variable process. In Tamminga (2014, 2016), I show that the monomorphemic (e.g., *ceiling*) and polymorphemic (e.g., *working*) context exhibit within-category, but not across-category, persistence, and argue that this is evidence that multiple variable processes are at play. In this paper, I aim to sidestep rather than illuminate these questions about the definition of the variable. Therefore, I exclude all monomorphemic observations and do not treat them as interruptions because I have already previously shown that they do not influence persistence in the much more frequent polymorphemic cases. On the other hand, in Tamminga (2014) I do find some puzzling evidence for persistence between the polymorphemic categories and *something/nothing*, a category that poses the additional problem of allowing additional variants. I therefore exclude the *something/nothing* category but conservatively treat *something* and *nothing* as interruptions. There is also one other special case, that of the phrase *going to*. I exclude instances of *gonna* from consideration entirely, but treat instances of *going to* that could have been produced as *gonna* as both exclusions and interruptions. Instances of *going to* that could not be realized with *gonna* (such as "I'm going to the store") are included normally.

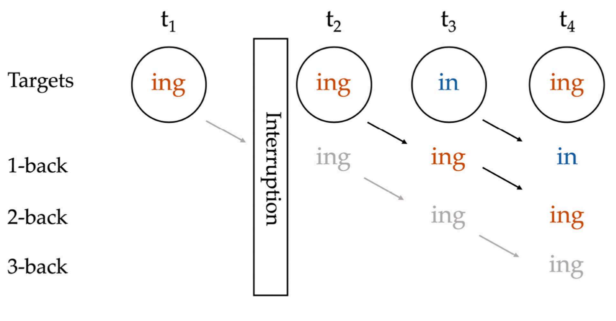


FIGURE 1 | Coding of a sequence with an interruption; grayed-out content reflects potential coding that is blocked by the interruption.

At each greater depth of N-prior sequence, some additional data is lost because of interlocutor and exclusion-based interruptions, and additionally the number of unique N-prior sequences increases. There is thus a tension between wishing to look at shorter N-prior sequences because there is more data and a simpler analysis, but also wishing to look at longer N-prior sequences because they provide a more refined view of the time-course of variable production. A 3-prior sequence seems to offer a good compromise between these goals in the particular data at hand, but I also look at the 2-prior and 4-prior sequences. The 2-prior sequence provides a simple starting point for reasoning about sequences of prior observations, and the 4-prior sequence makes it clear that the data at hand should not be stretched further. Overall, approximately the same general pattern arises at the 2-prior, 3-prior, and 4-prior levels, which provides some reassurance regarding the stability of the results.

2.3. Descriptive Analysis

I begin with an analysis of the subset of the verbal ING data for which the 2-prior sequence is intact (N = 3,071). For a depth of two prior observations, there are four unique prior token sequence options: /in-in-T/, /ing-in-T/, /in-ing-T/, /ing-ing-T/ (recall that T represents the linear position of the target). The first two sequences have /in/ as their immediately prior observation, and the last two sequences have /ing/ as their immediately prior observation, so a traditional persistence analysis would group together the first two sequences (as /in/-primed) and the last two sequences (as /ing/-primed). I calculated the /ing/ rate after each of these unique sequences. The results are in **Figure 2**. The unique 2-prior sequences are arranged on the x-axis, and the y-axis shows the probability of the /ing/ variant after each sequence. To help guide the visual interpretation at the expense of added redundancy, the graph is also faceted by how many /ing/ observations occurred in the 2-prior sequence, and the bars are color coded by the value of the 1-back variant. From **Figure 2**, it is immediately apparent that the /ing/ rate is higher for observations that had more instances of /ing/ in the 2-prior sequence: the /ing/ rate after two /in/ variants is 16% (N = 1,420), while the /ing/ rate after two /ing/ variants is 79% (N = 892). In the middle facet of the graph, we see an additional effect:

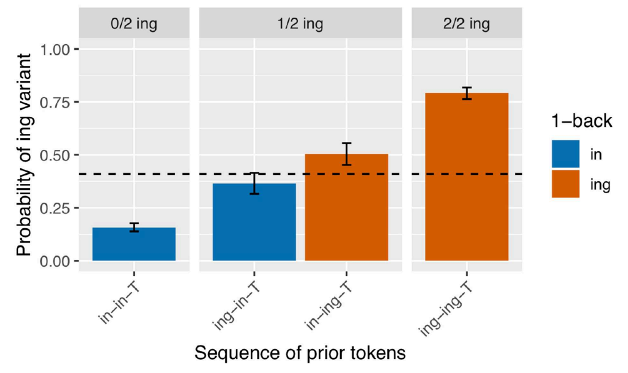


FIGURE 2 | Corpus probability of /ing/ variant by 2-prior sequence. Error bars are Clopper-Pearson binomial 95% confidence intervals.

when the 2-prior sequence contains one of each variant, the order they come in matters: the /ing/ rate is higher after an /in-ing-T/ sequence (50%, *N* = 375) than an /ing-in-T/ sequence (36%, *N* = 384). That there is a difference between the two blue bars and between the two red bars in **Figure 2** shows that the 2-prior sequence matters beyond supplying the immediate 1-back variant. But that there is a difference between the blue and red bars in the middle facet shows that there is an effect of the 1-back variant that goes beyond the total number of /ing/ observations preceding the target.

Next I turn to the subset of the data in which the full 3-prior sequence is intact (*N* = 2,334, so 737 observations removed from the 2-prior subset due to interruptions between the 2-back and 3-back positions). There are eight unique 3-prior sequences to consider, which I will not enumerate here but which can be found listed along the x-axis of **Figure 3**. **Figure 3** is set up in the same way as **Figure 2**: there is a bar representing the rate of /ing/ use for targets preceded by each of the unique 3-prior sequences, the facets represent the total number of /ing/ variants in the 3-prior sequence, and the color coding represents the 1-back variant. As before, we see a very strong effect at the far ends of the graph: the /ing/ rate after a sequence of three /in/ observations is 14% (*N* = 898) while the /ing/ rate after a sequence of three /ing/ observations is 83% (*N* = 540). In the 1/3 /ing/ facet, we see that the /ing/ rate is higher when the one /ing/ in the sequence is in the 1-back position (42%, *N* = 177) but that the order of the 2-back and 3-back positions does not make a large difference: the /ing/ rate is 26% after /ing-in-in-T/ (*N* = 192) and 28% after /in-ing-in-T/ (*N* = 142) sequences. In the 2/3 /ing/ facet, we see essentially the same thing: the /ing/ rate is depressed when the 1-back token was /in/ (47%, *N* = 132) but does not appear to differ between /ing-in-ing-T/ (61%, *N* = 108) and /in-ing-ing-T/ (60%, *N* = 145) sequences.

It should already be apparent from the token counts given in the discussion of the 3-prior sequence results that data sparsity will raise its head as a real problem in the 4-prior sequences, both because there are now 16 unique prior token sequences to subset by and because the total number of observations is down to 1804 after loss of an additional 530 observations due to interruptions between the 3-back and 4-back positions. However, even the smallest subset in this breakdown (/ing-in-ing-in-T/) still has 33 observations in it, so I will cautiously proceed. I will not break down all 16 /ing/ rates shown in **Figure 4** in the discussion here, but will instead make some general observations. With less data, the patterns are inevitably somewhat less clear, but there are a couple reasons to believe that the basic result here is consistent with the previous two clearer patterns. First, within each facet, every red bar is taller than every blue bar, and subsequently the average of the red bars is higher than the average of the blue bars across the three middle facets. This is consistent with the observation of an effect of the 1-back variant. Second, within the same-colored bars in each facet, the fluctuations we see are not consistent with plausible predictions from the sequence order. For instance, the /ing/ rate for /ing-in-in-in-T/ is higher than for /in-ing-in-in-T/ even though the latter has a more recent instance of /ing/ in the sequence. This suggests that the deeper-than-1-back order-based fluctuations seen here are random rather than systematic, and that if we had more data in each subset we would expect to see them level out to look more like **Figure 3**. Of course, the only way to confirm this would be to get more data, a non-trivial task.

2.4. Statistical Analysis

In the descriptive analyses just given in section 2.3, I took the following approach at each N-prior sequence depth. First, I calculated /ing/ rates conditioned on each unique N-prior sequence separately. Then, I proposed on the basis of those observed /ing/ rates that treating every unique prior token sequence as a distinct context was missing a generalization: that observed ING rates differ only based on how many /ing/ observations occurred in the prior sequence and what variant is in the 1-back position, not any additional information about the order of variants in the 2-back, 3-back, or 4-back positions. However, the descriptive analyses have not yet accounted for many factors that are known to affect variation in general or ING specifically, such as phonological context or speaker gender. They also do not account for the non-independence that results from different speakers (with different characteristic /ing/ rates) each contributing more than one token to the dataset (prior to the sequence formation process and associated exclusions for interruptions, the average number of observations per speaker is 34). I therefore turn to mixed-effects logistic regression to assess whether the observations I made based on the raw data reflect statistically significant differences that are robust to the inclusion of these other predictors.

The mixed-effects logistic regressions in this section were fit using the `lme4` package version 1.1-18 (Bates et al., 2015) in R version 3.5.1 (R Core Team, 2015). The dependent variable is the ING variant in each target observation, with 0 as /in/ and 1 as /ing/. The models include as fixed effects several known predictors of ING that are available in this dataset and were described in Section 2.1, namely lexical frequency, speech rate, preceding segment, following segment, and speaker gender. The lexical frequency measure (Lg10CD) comes from SUBTLEX (Brysbaert and New, 2009) already base-10 log-transformed, and speech rate is natural log transformed. These

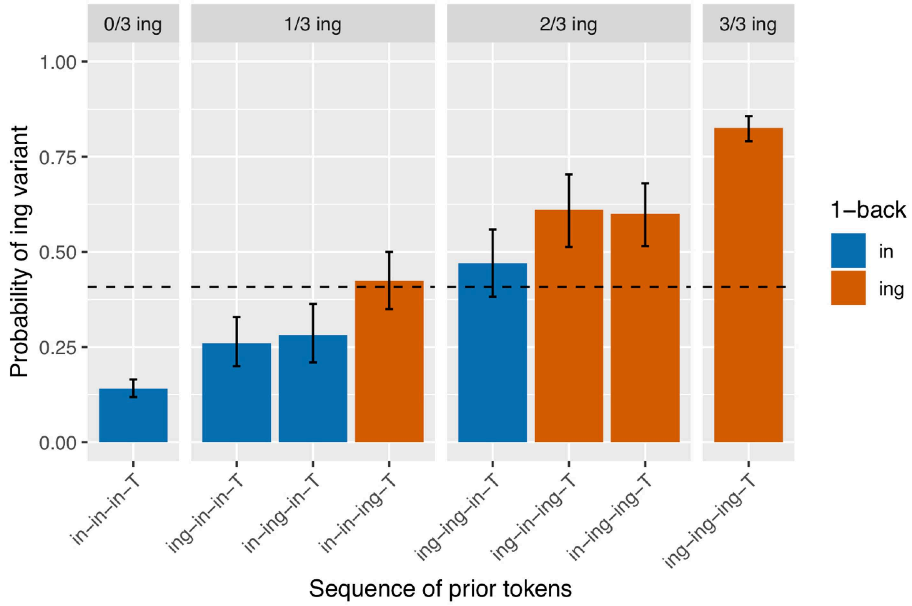


FIGURE 3 | Corpus probability of /ing/ variant by 3-prior sequence. Error bars are Clopper-Pearson binomial 95% confidence intervals.

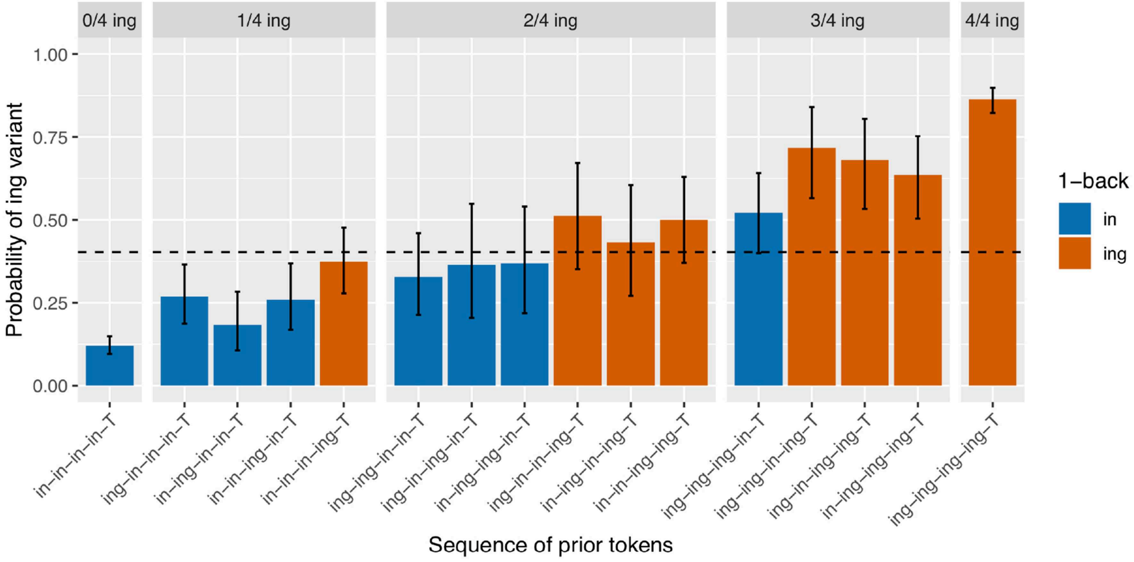


FIGURE 4 | Corpus probability of /ing/ variant by 4-prior sequence. Error bars are Clopper-Pearson binomial 95% confidence intervals.

continuous control predictors are then z-scored to center around their mean log value. The categorical control predictors (preceding/following phonological context and gender) are given a sum-coded (also known as deviation-coded) contrast scheme, so that the intercept in the regression is computed at the grand mean of their levels rather than a reference level. In addition to these fixed effects, each model also includes a speaker random intercept; equivalent models were fit with by-word random intercepts that were dropped because they captured little variance but made generating predicted values more complicated. The speaker random intercept is particularly important, as I discuss in Tamminga (2014), because the non-independence of observations from the same speaker can give rise to apparent "repetitiveness" effects without any true microtemporal clustering involved. Speaker clustering has not yet been controlled out in the mean rates shown in the figures above, so it is crucial to fit these models to account for that non-temporal source of apparent clustering.

I will focus on modeling the 3-prior subset of the data, attempting to capture the pattern seen in **Figure 3**, rather than modeling the 2-prior or 4-prior sequence analyses. I choose to focus on the 3-prior subset because the 2-prior sequences do not offer enough granularity to look at interesting sequence effects, while the 4-prior sequence analysis has so many prior token sequence conditions that it leaves us without enough data to get a confident probability estimate within each condition. I fit three models to the 3-prior data, which are intended to approximately map to the two-step approach I just recapped for the descriptive data analysis, with Model 1 representing the first step and Models 2 and 3 representing the second step and a refinement thereof. The fixed effects from Model 1 are given in **Table 1**. Model 1 includes a prior sequence predictor, with a separate level for each unique 3-prior sequence. The levels of this predictor are reverse difference coded, so each level is compared to the previous level. For example, the line in **Table 1** labeled "Prior seq. (ing.in.in.T - in.in.in.T)" represents the test of the difference between the probability of /ing/ in an /ing-in-in-T/ sequence and an /in-in-in-T/ sequence. The order of the levels is set to be the same as in **Figure 3**, so the coefficients in the model represent the difference between the height of each bar and the bar to the left of it (in log-odds). For example, the coefficient for "Prior seq. (ing.in.in.T - in.in.in.T)" maps to the estimated difference between the second blue bar from the left in **Figure 3** and the first one on the left.

The control predictors are all significant in the expected directions, which is good because they were selected to reflect only known influences on ING. When we turn to the critical predictor of prior sequence in this model, it is important to recall that the contrasts are set up so that each level is compared to the level preceding it. The order of the levels is the same as that in **Figure 3**: the levels are sorted first by their prior sequence /ing/ count, then by the 1-back position, then the 2-back position, reflecting a plausible expectation that more prior /ing/s might increase the /ing/ rate and, when the number of prior /ing/s is the same, those /ing/s might be expected to be more powerful if they are at a closer sequence position to the target. What we see is that the first three levels do not differ significantly from one another, but then /in-in-ing-T/ significantly favors /ing/ compared to /in-ing-in-T/ ($\beta = 0.78$, $p = 0.005$). The next level, /ing-ing-in-T/, does not differ significantly from /in-in-ing-T/, but it is significantly lower than /ing-in-ing-T/ ($\beta = 0.61$, $p = 0.045$). The /ing-in-ing-T/ level in turn does not differ significantly from /in-ing-ing-T/. But the final level, /ing-ing-ing-T/, does differ significantly from /in-ing-ing-T/ in favoring /ing/ ($\beta = 0.59$, $p = 0.017$). This set of hypothesis tests is consistent with my proposal that there is an influence of the 1-back variant but not deeper (that is, ($>$ 1)-back) order effects. The difference tests that are equivalent to the difference between each red bar with a blue bar next to it within a facet in **Figure 3**—that is, the jump up in /ing/ probability from 1-back = /in/ to 1-back = /ing/, when the prior /ing/ count is the same—show evidence that this 1-back effect is significant. The cases where the 1-back position and the prior /ing/ count are the same do not show evidence for a significant difference. Note that none of these predictors directly test the hypothesis of differences attributable to the prior /ing/ count alone. If there were no prior /ing/ count effect at all, we would expect the comparisons between levels where the 1-back value switches from /ing/ to /in/ but the prior /ing/ count goes up by 1 (as in the comparison between /ing-ing-in-T/ and /in-in-ing-T/ for example) to show a significant decrease in probability (essentially "resetting" back to the blue level instead of the red level). This is not the case. To directly test the idea that there are two things going on, prior /ing/ count and 1-back effect, I will need to fit a model containing those two predictors explicitly. The purpose of Model 1 here is in fact to argue that Model 1 is not the correct model: that in treating every 3-prior sequence as a unique context we are missing a generalization about how the real differences across those sequences can be captured by a pair of overlapping simpler predictors.

TABLE 1 | Model 1: Each 3-prior sequence compared to the previous 3-prior sequence.

	Estimate	z-value	Pr(>|z|)
Intercept	−0.38	−2.17	0.030
Control			
Speech rate	−0.22	−3.41	0.001
Lexical frequency	−0.45	−7.38	< 0.001
Preceding coronal	0.28	4.60	< 0.001
Following pause	0.35	4.73	< 0.001
Female speaker	0.49	2.99	0.003
Critical			
Prior seq. (ing.in.in.T - in.in.in.T)	0.12	0.56	0.575
Prior seq. (in.ing.in.T - ing.in.in.T)	0.05	0.19	0.847
Prior seq. (in.in.ing.T - in.ing.in.T)	0.78	2.81	0.005
Prior seq. (ing.ing.in.T - in.in.ing.T)	−0.27	−1.01	0.314
Prior seq. (ing.in.ing.T - ing.ing.in.T)	0.61	2.01	0.045
Prior seq. (in.ing.ing.T - ing.in.ing.T)	−0.16	−0.52	0.602
Prior seq. (ing.ing.ing.T - in.ing.ing.T)	0.59	2.40	0.017

Model 2, accordingly, is congruent with that proposal: instead of a single predictor with a different level for each prior token sequence, I include two predictors, one for /ing/ count in the prior sequence (the equivalent of the facets in **Figure 3**) and one for the 1-back variant (the equivalent of the bar colors in **Figure 3**). The prior /ing/ count is treated as a categorical predictor here, again using reverse difference coding for the contrasts. The results from this model are given in **Table 2**. There is a significant effect such that if the 1-back variant is /ing/, the target is more likely to be /ing/ ($\beta = 0.68$, $p < 0.001$). While the size of the coefficient is quite similar to the comparisons in Model 1 that amounted to a test of a 1-back effect while controlling prior /ing/ count (which were 0.78 and 0.61), pooling over all of the prior /ing/ count values approximately doubles the effect size (z). When we look at the prior /ing/ count predictor, we can see that the difference between 1 and 0 prior /ing/s is not significant but all other comparisons between levels are. This is consistent with what we saw in Model 1 with the lack of difference between the first two levels of the prior sequence predictor.

Model 3 reflects a refinement of Model 2 but keeps the basic premise of the model. The only difference between Model

2 and Model 3 is that Model 3 treats prior /ing/ count as a continuous numeric predictor instead of a categorical predictor. In one sense this is not the correct thing to do: an integer count value is a different sort of thing than a continuous number, and the only options for prior /ing/ count values are integers. However, what it reflects in this model is the premise that what we're trying to capture with the prior /ing/ count predictor is something like "how /ing/-ful is the speaker's overall recent prior experience," and we only have a coarse-grained measure of what is underlyingly a continuous measure. In theory we might want to look at something like a weighted moving average over a larger window to get a more truly continuous measure of "how /ing/-ful is the speaker's overall recent prior experience." The reason I do not undertake such an analysis is that the problem of interlocutor interruptions makes it difficult to go very far back. In any case, **Table 3** presents the results of Model 3. It shows that the 1-back estimate is stable but now the linear prior /ing/ count predictor has a larger effect size and much smaller *p*-value than any of the corresponding prior /ing/ count values in Model 2.

The three models I have fit here are not nested, and therefore cannot appropriately be compared formally with log-likelihood tests. However, various model criteria might support an informal comparison of the models. Each model is simpler than the last in terms of degrees of freedom (Model 1 d.f. = 14, Model 2 d.f. = 11, Model 3 d.f. = 9). As a result, the log likelihood inevitably goes up, but only slightly: the log likelihoods of the three models are −1058.2, −1058.8, and −1059.4, respectively. Meanwhile, the AIC and BIC measures, which penalize extra parameters, go down from Model 1 (AIC = 2144.5, BIC = 2225.1) to Model 2 (AIC = 2139.5, BIC = 2202.8) and from Model 2 to Model 3 (AIC = 2136.8, BIC = 2188.6). These criteria are in line with the view that Model 3 is the simplest and strongest model of the prior sequence effects in this data.

Figure 5 shows a data visualization that is equivalent to the observed data visualization in **Figure 3** but instead represents the predicted probabilities from Model 3 for a particular male speaker (PNC PH06-2-2) whose mean /ing/ rate is near the dataset grand mean, for a token that neither follows a coronal nor precedes a pause and has a scaled log vowels per second of 0 and a scaled Lg10CD value of 0. This illustrates that this model is producing predictions that are a good match for the empirical patterns we saw in section 2.3—these patterns remain when we control for speech rate, frequency, phonological context, speaker gender, and speaker identity clustering.

3. PRIOR SEQUENCES IN SIMULATED DATA

The empirical data in section 2.3 showed the same pattern at three lengths of N-prior sequence: the probability of /ing/ at a target is affected by both the total number of /ing/ instances in the N-prior sequence and the variant used at the 1-back position (that is, the token that would normally be treated as the prime), without evidence to suggest that it is influenced by the order of prior observations at an N-back position of N greater than 1. The statistical modeling in section 2.4 supported that interpretation of the data while controlling for other known predictors of ING. But what does this result actually tell us about the source of persistence? In this section I aim to show that this type of analysis can move us toward an answer on a problem that has seemed intractable for some time.

In this section I use a series of simple Bernoulli process simulations to explore the potential processes generating different patterns of target probabilities based on prior token sequences. It should be emphasized that this is a preliminary tour through what I believe could become a fruitful area of research more broadly. The use of computational simulations in sociolinguistics is not new, but most simulations are simulations of communities, such as agent-based models of the spread of sound change through a population over generations. The simulations I use here are focused on a microtemporal level and are conceptually very simple: I model the production of variation essentially as strings of coin flips at different probabilities, then analyze the generated data in the same way as I analyzed the corpus ING data. I compare the output of different simulated models to the corpus results from the previous section as a way of investigating the plausibility of different processes having generated the data. I particularly pay attention to the ways

TABLE 2 | Model 2: Categorical prior /ing/ count and 1-back.

	Estimate	z-value	Pr(>|z|)
Intercept	−0.68	−3.62	<0.001
Control			
Speech rate	−0.22	−3.46	0.001
Lexical frequency	−0.45	−7.35	<0.001
Preceding coronal	0.28	4.58	<0.001
Following pause	0.35	4.72	<0.001
Female speaker	0.49	2.99	0.003
Critical			
1-back /ing/	0.68	4.07	<0.001
Prior /ing/ count (1-0)	0.20	1.12	0.262
Prior /ing/ count (2-1)	0.38	2.15	0.032
Prior /ing/ count (3-2)	0.47	2.32	0.020

TABLE 3 | Model 3: Continuous prior /ing/ count and 1-back.

	Estimate	z-value	Pr(>|z|)
Intercept	−1.19	−6.38	<0.001
Control			
Speech rate	−0.22	−3.50	<0.001
Lexical frequency	−0.45	−7.31	<0.001
Preceding coronal	0.28	4.60	<0.001
Following pause	0.35	4.69	<0.001
Female speaker	0.49	3.03	0.002
Critical			
1-back /ing/	0.67	4.06	<0.001
Prior /ing/ count	0.35	3.79	<0.001

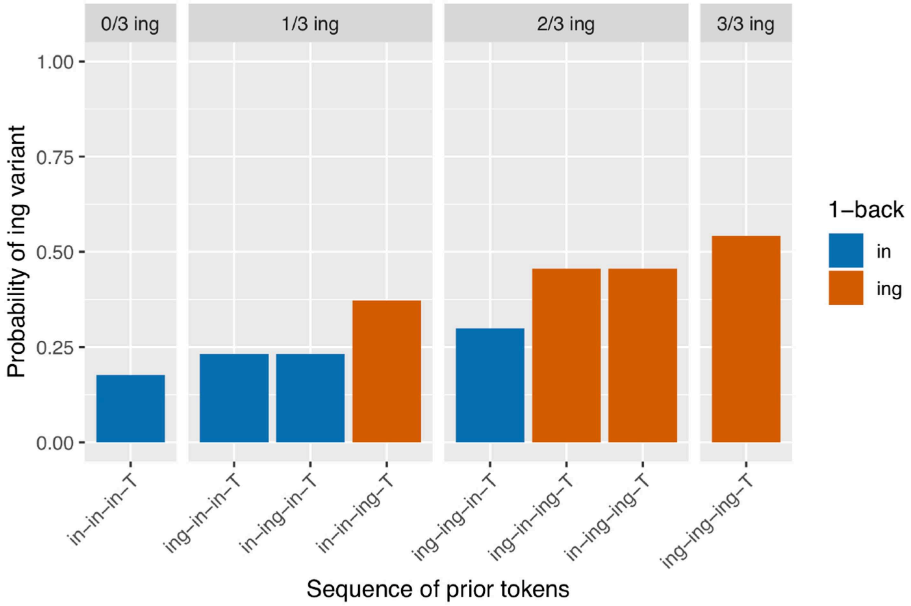


FIGURE 5 | Predicted values from Model 3 (male speaker PH06-2-2 with observed /ing/ probability = 0.4, non-pre-coronal, non-post-pausal, scaled log vowels per second = 0, scaled Lg10CD = 0).

in which the predictions from models of baseline deflection and models of sequential dependence are dissociated under various conditions. This is of interest because it motivates the study of multiple token sequences in contrast to the usual persistence approach (looking at only one prior token) that does not distinguish between baseline deflection and sequential dependence. While I will not be able to conduct an exhaustive search of the many-dimensional parameter space opened up by these models, my preliminary explorations here will suggest that a model combining both a sequential dependence mechanism and a baseline deflection mechanism produces patterns that most closely and consistently resemble the results of the corpus data analysis in section 2.3.

3.1. Simulation Preliminaries

For clarity of exposition with a sociolinguistic audience in mind, I will discuss the models here as if they involved speakers producing the ING variable: for instance, I will describe a Bernoulli trial[9] with an outcome of 1 as an instance of the /ing/ variant. I will also present visualizations of the model outputs using this framing around ING, making the graphs directly visually comparable to the graphs in section 2.3. It should, of course, be borne in mind that everything happening in these simulations is merely lists of probabilities and 0s and 1s; nothing about them is specific to ING (or to sociolinguistic variation, or, indeed, to linguistic behavior).

Each simulation involves the same set of simulated "speakers," whose identity is tracked during each run of the simulation. Each speaker has some baseline probability of producing the /ing/ variant (vs. the /in/ variant). These baseline probabilities are taken from the observed corpus data so that the overall distribution of speakers and their linguistic behavior resembles that of the real data. In the corpus 3-prior dataset, there are 118 speakers who each produce on average 34 observations. Of these, 17 speakers end up contributing only /ing/ or only /in/ outcomes to the 3-prior data, but only because of exclusions: none of these are speakers whose ING behavior is categorical in the larger data set. However, in the interest of avoiding simulated speakers with categorical baselines, I exclude these 17 speakers in order to end up with 101 simulated speakers with non-categorical baselines. The distribution of by-speaker baseline /ing/ probabilities is shown in **Figure 6**. Each of the simulated speakers will produce an ordered string of 20 "ING tokens" (Bernoulli trials) with the speaker's /ing/ probability as the outcome probability of each trial. Since the first three trials from each speaker are excluded from analysis because they do not have enough previous trials, each speaker contributes 17 observations to the simulated data set, resulting in a total of 1717 observations in each simulated data set (compared to 2300 in the observed data at 3-prior depth). I calculate the observed proportion of 0s and 1s conditioned on each preceding trial sequence, then store these values. The entire run is then repeated 500 times and the distribution of results from those runs is presented graphically. I also fit a linear mixed effects regression to each simulation run, with predictors equivalent to the critical predictors from Model 3 from the corpus data analysis plus the speaker random effect (the control predictors in Model 3 are not relevant for the simulated data). I extract the 1-back

[9] A Bernoulli trial is simply a random variable with only two possible outcomes, sometimes treated as "success" and "failure." The probability of success and probability of failure add up to 100%. A familiar example of a Bernoulli trial is a coin flip.

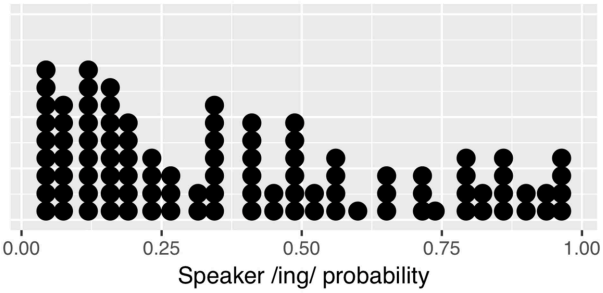


FIGURE 6 | Observed by-speaker probabilities from corpus data, used for simulated speaker baselines.

and prior /ing/ count predictor z-values (effect sizes) and p-values from each run over the course of the 500 runs in order to find out how often each simulation produces statistically significant effects aligned with the corpus results.

The series of simulations that I will compare across the following subsections is built up as follows. The first simulation, in section 3.2, contains no microtemporal clustering: I call this the null simulation. Each subsequent simulation has some source of microtemporal clustering added in. In the sequential dependence simulations in section 3.3, the built-in clustering mechanism that is added to the null simulation is that the outcome of each trial affects the outcome probability for the next trial. In the baseline deflection simulations in section 3.4, a different built-in clustering mechanism is added to the null simulation: each speaker has two or more states with distinct target probabilities that are above and below the speaker's characteristic probability. These create the possibility of baseline deflection as the speaker moves between different states and thus different target probabilities; a Markov chain model generates the sequences of states that the speakers move through. Finally, in section 3.5, both of these distinct clustering mechanisms are included in the simulation at the same time. In all simulations, the data is generated by sampling the binomial distribution randomly at each trial (at the specified probabilities) using the `binom` package in R.

3.2. The "Null" Simulation: No Microtemporal Clustering

The first thing I do is show what the N-prior sequence effects look like in data that has speaker clustering (speakers differ in their characteristic rates) but no form of microtemporal clustering (that is, neither sequential dependence nor baseline deflection, with no intraspeaker probability fluctuation). I call this the "null" simulation because of the lack of critical clustering structure. This simulation is important because it would be easy to mistake speaker clustering for within-speaker temporal structure. This will also be a starting point for the creation of various microtemporally structured probability patterns that I will use in the subsequent simulations.

The speaker baselines in the null simulation are as just discussed in section 3.1 and shown in **Figure 6**. The results of the null simulation are shown in **Figure 7**. What is immediately apparent is that the effect of the prior /ing/ count seen in section 2.3 arises from speaker clustering without any within-speaker microtemporal structure. This makes sense: without controlling for speaker clustering, a target preceded by three /ing/ outcomes is more likely to be a target from a high-/ing/ speaker and therefore more likely to itself have an /ing/ outcome. While there would be an apparent 1-back effect if we looked only at the 1-back prior token depth (the red boxes are on average higher than the blue boxes), we do not see any 1-back effect beyond that generated by the prior /ing/ count, which is also as expected. The regression results from the simulations confirm that the 1-back effect is not present (a significant positive effect on 1.8% of trials and a significant negative effect on 2.8% of trials).

In theory, including random speaker intercepts in a linear mixed effects model of each simulation's data should eliminate the visually-apparent /ing/ count effect. The statistical model values show that actually the models end up somewhat anti-conservative: there is a significant positive effect of prior /ing/ count on 11.4% of runs. Because the structure of the model does not include any possible true microtemporal source of this effect, we can be confident that these findings actually arise from incompletely controlled speaker clustering [10]. This should be kept in mind when interpreting the other models; I will compare the observed number of significant prior /ing/ count effects to this rate [11].

3.3. Simulating Sequential Dependence

I now build on the null simulation by adding the first candidate source of within-speaker microtemporal structure: sequential dependence. This simulation is identical to the previous one except that, within each speaker, the outcome probability of each Bernoulli trial is slightly influenced by the outcome of the previous trial. I set the probability adjustment to 0.05: if the prior outcome was a 1, I add 0.05 (out of 1) to the target probability, and if the prior outcome was a 0, I subtract 0.05 from the target probability. The probability adjustment is always done to the speaker's base probability, so the probabilities don't snowball and go out of bounds. Notice that this is equivalent to each speaker having two states with different /ing/ probabilities, with the state they are in on each trial determined by the ING outcome of the previous trial. Any number of more sophisticated adjustments to the baseline could be used to generate the exact /ing/ probabilities for these states; the ± 0.05 adjustment is simple and transparent but is not intended to involve any substantive claim about how these probabilities are or should be adjusted.

The results we see in **Figure 8** bear a resemblance to the observed data in **Figure 3**. We see what looks like the prior /ing/ count effect, although the null simulation made it clear that this can derive from speaker-level clustering. We also see

[10] A "true" null simulation would be one that simply contains 1717 Bernoulli trials at a single probability, which should produce a spurious prior /ing/ count effect only 5% of the time.

[11] Of course, the empirical data should also be reassessed in light of this finding, but because the prior /ing/ count *p*-value from Model 3 is very low, I will continue with the assumption that this effect is unlikely to be due to chance even with the elevated probability of a spurious result.

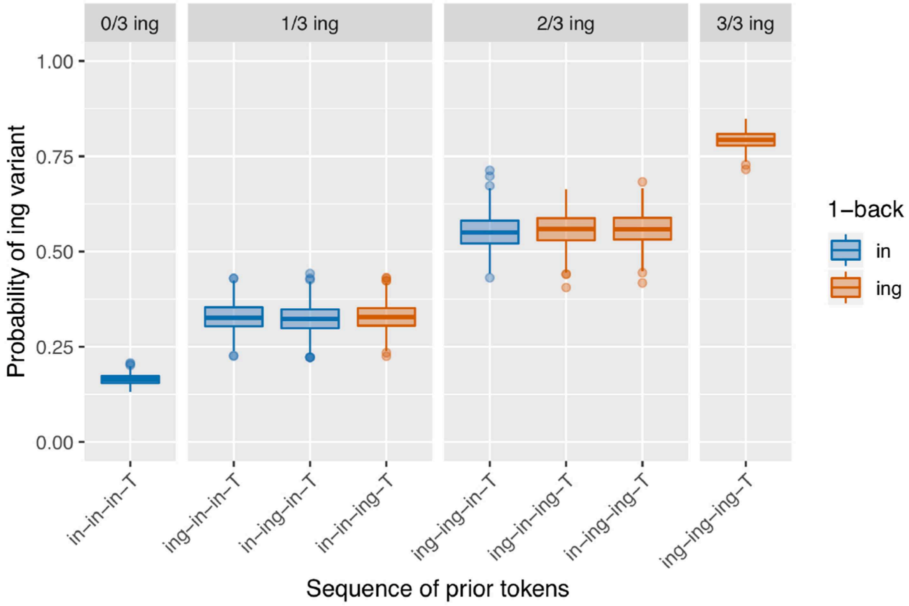


FIGURE 7 | Simulation with speaker baseline differences but no built-in microtemporal clustering.

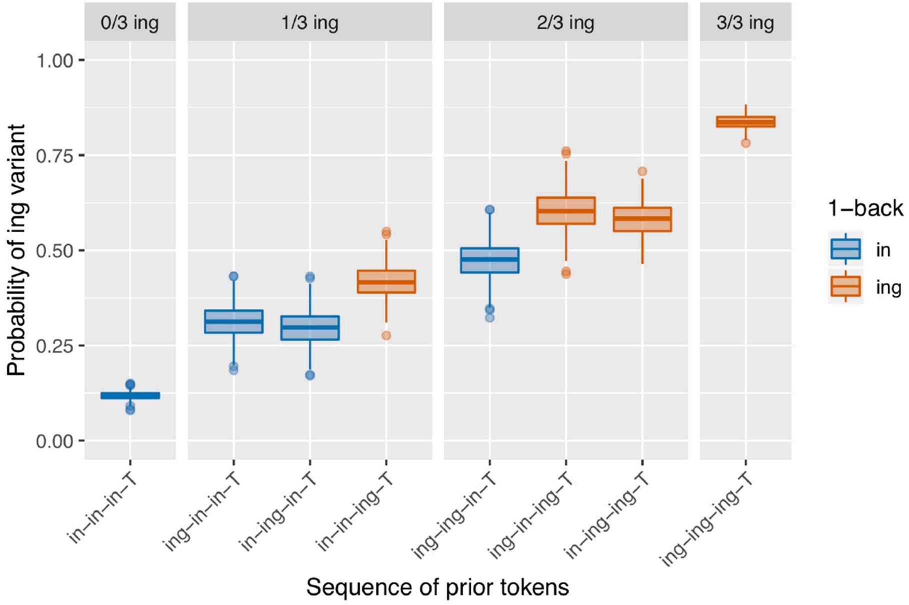


FIGURE 8 | Results of 500 runs of a sequential dependence simulation with a 0.05 boost.

an effect where the red boxes are higher than the blue boxes within each facet: the only-1-back effect. This model produces a significant positive 1-back effect on 73.4% of runs, but a significant positive prior /ing/ count effect only 8.6% of the time—the latter being slightly lower than the false positive rate in the null simulation. In other words, all of the apparent /ing/ count effect here is attributable to the speaker rather than temporal clustering. Interestingly, there is also a small difference between the /ing-in-in-T/ and /in-ing-in-T/ conditions in the 1/3 ing facet, and between the /ing-in-ing-T/ and /in-ing-ing-T/ conditions in the 2/3 ing facet. These differences result from small biases in which types of speakers produce which prior token sequences [12]. Consider the 2/3 /ing/ sequences. If a speaker has a low /ing/ baseline probability, they are slightly

[12] Thanks to Dan Lassiter for identifying the source of these differences.

more likely to produce an /ing/ after another /ing/ (as in /in-ing-ing-T/ due to the facilitating effect of the first /ing/) but less likely to spontaneously produce /ing/ twice apart from that facilitating influence, as in /ing-in-ing-T/. In contrast, it is somewhat "easier" for a high-/ing/ speaker to produce the two /ing/s spontaneously. As a result, /ing-in-ing-T/ prior sequences are slightly more likely to come from high-/ing/ speakers, and subsequently slightly more likely to result in an /ing/ outcome.

3.4. Simulating Baseline Deflection

In the next set of simulations, I investigate baseline deflection instead of sequential dependence. I remove from the simulations the mechanism of adjusting the target probability based on the prior outcome. Instead, I give each speaker two target probabilities that average to the same characteristic probability as they had in the previous simulations, when possible. Specifically, I add and subtract 0.3 from the baseline, so for example a speaker with an overall baseline of 0.4 will have a state A /ing/ probability of 0.1 and a State B /ing/ probability of 0.7. When this calculation would put the probabilities outside of the 0 to 1 range, I replace the value with 0 or 1 accordingly—so, speakers can have a categorical behavior in one of their two states. The speaker then switches back and forth between states A and B over stretches of trials.

The state-switching behavior in the simulation is generated stochastically using a Markov process: each state has a transition probability reflecting the likelihood that the process will switch to the other state for the next trial, but there is no further time dependence. I use symmetrical transition probabilities throughout the simulations I present here (so the probability of switching from A to B is the same as the probability of switching from B to A) but will present several different transition probabilities reflecting different degrees of state "stickiness." The use of the Markov process to generate the state switches is not intended as a claim that this kind of state switching is actually generated stochastically. On the contrary: I expect that changes in state would reflect responses to changes in the real world context where the speech is taking place, such as changes in topic, context, or interlocutor, or changes in the speaker's internal state, such as shifts in stance, attitude, or attention. From the perspective of the analyst, however, such contextual changes are unpredictable and therefore can be modeled as a stochastic process [13]. Once the sequence of states has been determined, there is a Bernoulli trial with the probability of success equal to the output probability at each trial's predetermined state, which produces the /ing/ or /in/ variant as in the previous simulations. The idea is to produce a model capturing the intuition that when two trials are closer together they are more likely to be in the same state, and therefore more likely to have the same outcome. The most important property of the model is simply that the state sequences are generated independent of the outcomes at each trial.

This approach to the simulation of baseline deflection offers different parameters that could be adjusted to generate a very wide range of possible outcomes. Here I present versions of the simulation at four different between-state transition probabilities. I do not change any other parameters: I hold the number of states (two) and the size of the difference between them for each speaker constant and do not allow for one state to be stickier than the other or for the stickiness of states to change over time.

When the transition probability is low, so the states are quite sticky, the result is a pattern that reflects the continuous effect of a prior token sequence such that the more prior /ing/s there are, and the closer in the sequence they are to the target, the higher the observed /ing/ rate in the target will be. This is shown in **Figure 9** for a model where the transition probability out of both states is 10%. I call this a continuous-N-back effect, in contrast to an only-1-back effect. In the regression models extracted over the runs of the simulation, this simulation produces a significant positive /ing/ count effect on 99.8% of runs, and a significant positive 1-back effect on 71% of runs. This seems promising, but recall that the model is not actually set up to detect a difference between a continuous-N-back effect and an only-1-back effect; visual inspection of the output in **Figure 9** suggests that this is a somewhat different pattern than what we see in the corpus data. In a model where the transition probability is 50% for both states, so speakers are equally likely to stay in their current state or switch to the other state, then both the 1-back and prior /ing/ count effects are lost: there is a significant positive /ing/ count effect on 10.8% of runs, again comparable to the null rate, and a significant positive 1-back effect on 1.4% of runs. The output of the model is not shown here but is visually identical to that of the null model.

It is also possible to get a result that looks like the 1-back result from the sequential dependence model. This arises when the transition probability for both states is just shy of 50%, so a speaker is a little more likely to stay in their current state than not: **Figure 10** shows the results when the transition probability is 40%. This model produces a significant positive 1-back effect on 36.6% of runs, which is not trivial but also not as good as the sequential dependence model where 73.4% of runs produce a 1-back effect. Like the sequential dependence model, though, this simulation mostly loses the significant prior /ing/ count effect, producing a significant positive /ing/ count effect on only 17.8% of runs, not a very big improvement over the 11.4% positive results in the null simulation.

Interestingly, these simulations are also able to reverse the direction of at least the 1-back pattern. **Figure 11** shows that as soon as the transition probability in each state is over 50%, the direction of the 1-back effect reverses, so that at each value of the prior token count, the contexts where the prior /ing/s were further away have the higher /ing/ probability, which is not as we would generally expect given the usual persistence pattern. The statistical models confirm this reversal: on 67.6% of runs of this simulation there is a significant negative 1-back effect. This reversal reflects the fact that when the transition probability is over 50%, two sequentially adjacent tokens are actually *less* likely to occur in the same state, rather than more likely, because from token to token the state is more likely to switch than to stay the same. This highlights that the argument in favor of baseline deflection as a source of repetitiveness does contain some assumptions about the time course of baseline deflection,

[13] Thanks to Kory Johnson for this suggestion and for proposing the use of Markov processes for this purpose.

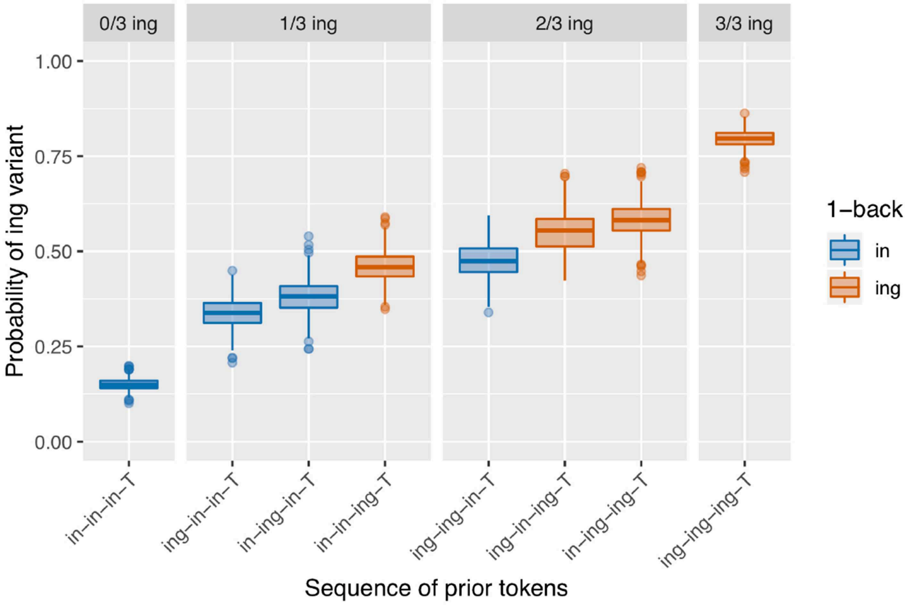


FIGURE 9 | Results of 500 runs of a baseline deflection model with between-state transition probability of 0.1.

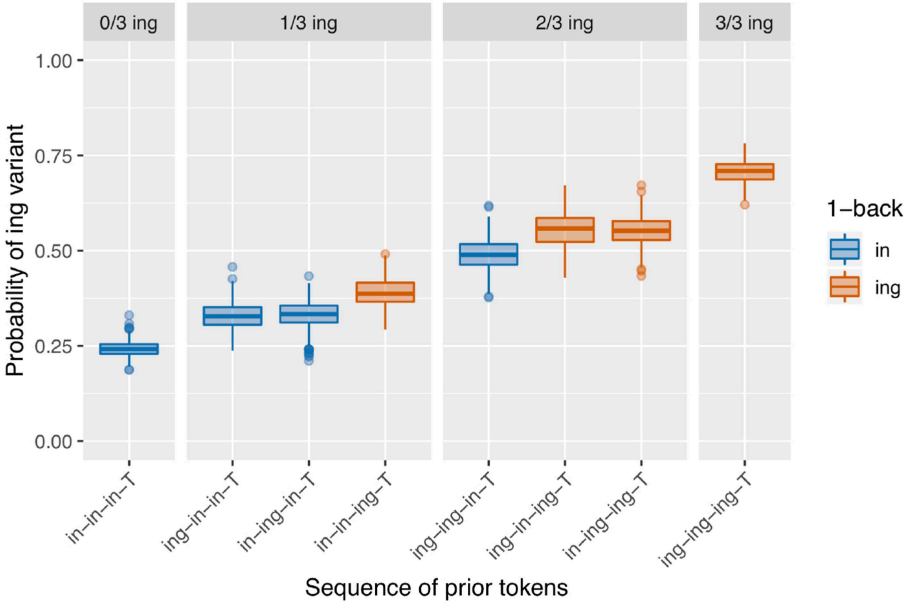


FIGURE 10 | Results of 500 runs of a baseline deflection model with between-state transition probability of 0.4.

namely that the window over which the baseline might shift is sufficiently wide that in fact two tokens occurring sequentially are more likely to be produced in the same window than not. It is also worth noting that there is an attested pattern of anti-persistence in the literature, which Szmrecsanyi (2006) terms the *horror aequi* effect. This particular simulation gives us one way of understanding how such an effect could arise.

3.5. Combining Sequential Dependence and Baseline Deflection

Both of the simulation types discussed so far have drawbacks in terms of the likelihood that their microtemporal clustering model might have produced the corpus ING data discussed in section 2.3. The sequential dependence model nicely produces an only-1-back effect reminiscent of the distinct pattern seen in the corpus data, but produces a prior /ing/ count effect only at chance

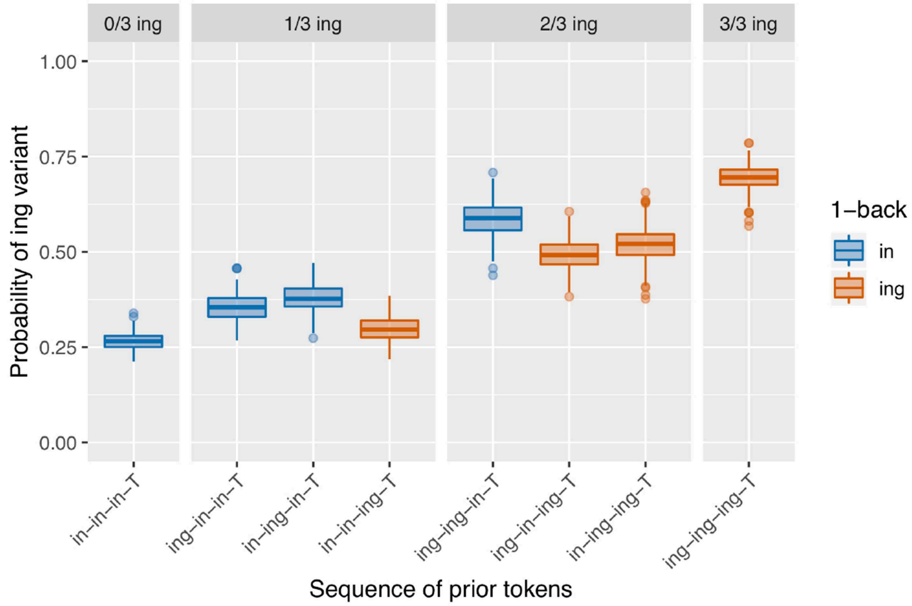


FIGURE 11 | Results of 500 runs of a baseline deflection model with between-state transition probability of 0.6.

rates. The baseline deflection models can clearly produce a wide range of patterns. But in the case where a baseline deflection model does consistently give rise to the desired prior /ing/ count effect (the version with the lowest transition probability), it also produces a continuous-1-back pattern rather than an only-1-back pattern.

There are two model classes under consideration here, and two empirical effects we desire to produce with the models. It seems that each model is better suited to producing one of the empirical effects: most versions of the baseline deflection models produce an /ing/ count effect, and the sequential dependence model produces an only-1-back effect. An appealing next step, then, is to combine the models to create a simulation that has both sequential dependence and baseline deflection built in. In this simulation, the state-shifting behavior is first generated using a Markov process as in the baseline deflection models; then the coin-flipping procedure takes place with the sequential dependence boosting behavior built in. The results of a set of simulations of this type with transition probability of 0.1 (as in the baseline deflection model of **Figure 9**) and a boost of 0.05 (as in the sequential dependence model of **Figure 8**) are shown in **Figure 12**.

This set of simulations now has several desirable features. The basic pattern of results shown in the graph more closely resembles an only-1-back effect than a continuous-1-back effect, making it an improvement over the component baseline deflection model alone; this is achieved through the inclusion of the sequential dependence boost. In terms of the model fit, we get a significant /ing/ count term on 99.2% of runs and a significant 1-back term on 99.6% of runs. By combining these two sources of microtemporal clustering into a single model—in a way that is consistent with the existence of multiple independently motivated phenomena that we expect to shape linguistic behavior in speech—we are able to more consistently arrive at an outcome that resembles the corpus data.

4. DISCUSSION

The sizable corpus sociolinguistic literature on persistence has typically asked how a single prior instance of a variable affects the outcome in a target instance of the same variable. In the first part of this paper, I extended this view of persistence to ask what effect sequences of multiple prior tokens have on the outcome of a target token. The descriptive results in section 2.3 indicate that this analysis of sociolinguistic sequences can reveal additional microtemporal structure that is not visible when we look only at a single prior token. More specifically, there are two aspects of the corpus ING results that are of interest and would not be detectable with the 1-back information only. First, there is a cumulative effect of how many /ing/ tokens occur in the prior token sequence, regardless of their position. This effect goes beyond the clustering we expect merely from differing speaker baselines. Second, there is a distinct effect of what variant occurred in the 1-back position. If we look only at the previous token, we would not be able to see either effect: we could not tell the difference between 1/3 and 2/3 of the prior tokens being /ing/ if we had only one token, nor would we be able to tell that the order of previous tokens is irrelevant beyond the 1-back position.

In the second part of the paper, I have suggested that this enriched view of the microtemporal structure of sociolinguistic repetitiveness can bring new evidence to a longstanding debate about the nature of that repetitiveness. The observation of

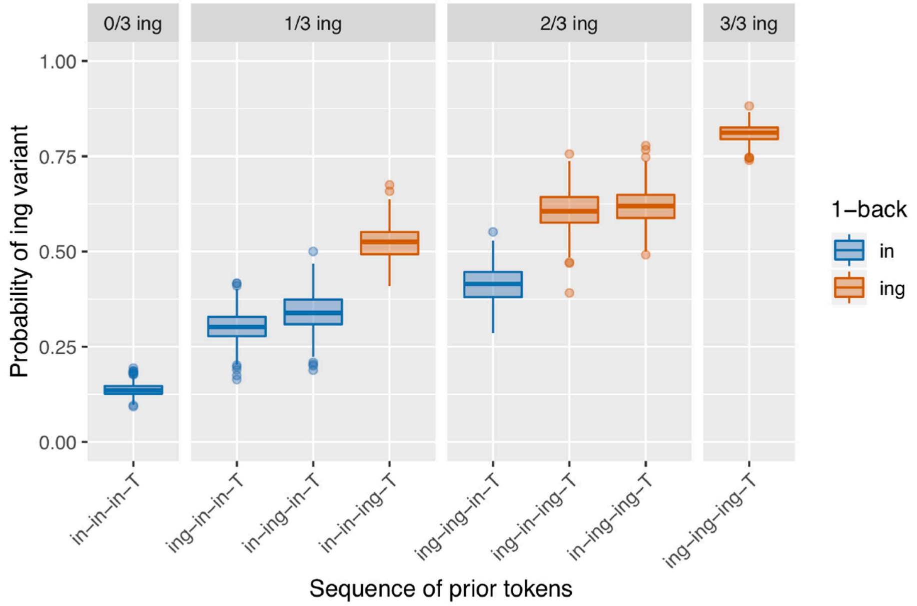


FIGURE 12 | Model with both baseline deflection (transition probability = 0.1) and sequential dependence (boost = 0.05).

persistence in corpus data has often been interpreted as reflecting sequential dependence, where the outcome of a prior instance of the variable directly influences the target outcome. On the other hand, it is often objected that persistence might arise as a result of baseline deflection, where sequential tokens are more likely to occur under similar contextual circumstances and therefore more likely to have the same outcome. To clarify what these two types of microtemporal clustering predict, I built a number of simulations in which sociolinguistic variation between /ing/ and /in/ is modeled using Bernoulli processes. In these simulations, sequential dependence is modeled by allowing the outcome of one Bernoulli trial to adjust the outcome probability on the next Bernoulli trial, while baseline deflection is modeled by creating pre-established sequences of states with different outcome probabilities but then not making reference to the actual *outcomes* across trials.

The sequential dependence model produces one of the two central effects of interest in the empirical data, the only-1-back pattern (seen in **Figure 8**). From a mechanical point of view, this can be understood straightforwardly: the sequential dependence models were built such that the target trial is only given information about the outcome of the immediately prior trial, not of previous trials. Of course, nothing would prevent us from building a sequential dependence model that adjusts the target trial probability based on the outcome information from several previous trials. The corpus result, then, is not trivial; the usefulness of a sequential dependence model that only tracks a single prior token suggests that it may be worthwhile to investigate comparable real-world processes that operate over long distances in terms of time yet a limited window in terms of prior instances of the linguistic variable. A downside of the sequential dependence model is that it does not reliably produce the /ing/ count effect. It is possible to build a baseline deflection model that mimics the output of this sequential dependence model (as in **Figure 10**), but such a model ends up with the same drawback as the sequential dependence model in that it also does not reliably produce the /ing/ count effect. On the other hand, a baseline deflection model with a relatively low between-states transition probability of 0.1 has the advantage of almost always producing a significant /ing/ count effect as desired. However, it does not produce the same kind of separation between 1-back (and only 1-back) conditions as the corpus data exhibits. Instead, it produces a continuous effect of recent /ing/ tokens: the more /ing/s and the closer those /ings/ in the prior token sequence, the greater the likelihood of /ing/ in the target (as seen in **Figure 9**). While we might have expected such a continuous-N-back effect on intuitive grounds, it does not actually accord with the pattern seen in the corpus data. In section 3.5, I showed that combining the sequential dependence and baseline deflection clustering mechanisms into a single model produces a surface pattern that is a near match for the corpus data, as well as nearly-always significant critical main effects from the regression models.

That the combined simulation seems to most successfully match the corpus data is an appealing result because we have independent evidence for the real-world phenomena that might produce both types of microtemporal clustering. As I discussed in section 1.1, there are multiple candidate phenomena that might give rise to each of the two types of microtemporal clustering under consideration here. Priming is the most commonly appealed-to phenomenon generating sequential dependence, but other sources of true sequential dependence are possible. Style shifting, broadly construed, is the most frequently suggested phenomenon that could give rise to co-occurrence through baseline deflection. To reiterate the point in section 1.1, nothing in this paper should be taken as evidence for or against particular mappings of clustering types to real-world interpretations. However, the fact that phenomena that could produce both clustering types unquestionably *exist* means that a model in which multiple phenomena are at play is an entirely plausible one. For example, were we to think that baseline deflection arises from contextual style-shifting while sequential dependence arises from priming of a recently-used linguistic option, we might find it entirely unsurprising that speakers are both style-shifting and exhibiting priming at the same time: there is plenty of evidence for the existence of both style-shifting and priming in human linguistic behavior. Indeed, to conclude that one of those phenomena was not at play might be even more surprising. The same logic applies to other possible interpretations of the sources of microtemporal clustering; the current study has nothing to say about where sequential dependence and baseline deflection come from, although conceivably some outgrowth of this approach could be used to probe for more precise quantitative properties of priming and style-shifting in future work.

Of course, the analyses and results of this paper are far from conclusive; they are best treated as a promising methodological demonstration inviting further research. One possibility that should be kept in mind is that the particular properties of the corpus results themselves could have occurred by chance. I have explored the simulations with a view to identifying a model that could plausibly have generated the corpus results as observed. But given the role of chance as well as possible uncontrolled factors in conversational speech data, one possibility is that the corpus results themselves are a chance output of a model like one of the models I have deemed less successful. Even if the pattern of results seen here is not due to chance, it might still be true that the pattern reflects something specific about the particular conversational interactions in the PNC data, or something unique to Philadelphia English, or something about the ING variable itself. We should be cautious to not reify or over-interpret the "prior /ing/ count" and "only-1-back" effects as I have described them here. The basic persistence effect has been found repeatedly across many different studies and therefore is seen as demanding a relatively general explanation; no deep investment in general explanations of these longer sequence effects should be made unless they can also be established as more generally recurring properties of sociolinguistic sequences. The most important step toward building confidence in this pattern of results will be to repeat the analysis on other ING data sets, other English variables besides ING, and ideally other languages entirely.

There are also many possible analyses that this paper has not undertaken. My preliminary explorations of the simulations have barely broached the many-dimensional parameter space afforded even by the simple models used here. Furthermore, the models could be enriched in many ways. While it would probably not be useful to simulate all of the possible details of ING variation simultaneously, one particular factor that has not played a role in any of the analyses thus far is the amount of time that elapses between each token. In previous work I have shown that the decay of ING persistence is very slow (Tamminga, 2014), which suggests that decay is unlikely to play a major modulating role in the effects we see when we abstract away from the exact duration of the time between a prior token and a target. An additional practical consideration in omitting temporal lag as a factor in the corpus analysis is that it is not, at first glance, obvious how best to combine the different prior token sequences with all of the possible decay relationships between them. However, future work might explore ways of integrating a continuous time dimension into the analysis of prior token sequences.

The goal of this paper was to show that there is value in the study of sociolinguistic sequences and the microtemporal structure they reveal. Sequential dependence and baseline deflection seemed inextricably intertwined in the 1-prior view, and indeed every single simulation in section 3 produces an overall difference between 1-prior conditions that would be counted as a finding of persistence under traditional quantitative approaches to persistence. Through the simulations, though, we learned that a longer time window can give us a more nuanced picture of what speaker repetitiveness looks like, with baseline deflection and sequential dependence producing outcomes that can be seen to be different when we look at longer prior sequences. We have already made much progress through the study of persistence at the 1-prior depth; as Szmrecsanyi concludes, "persistence is actually sufficiently patterned and predictable to help us understand better the linguistic choices that speakers make" (Szmrecsanyi, 2006, p. 6). The combined corpus analysis and simulations here suggest that this sentiment is as true of longer sequences as it is of prime–target pairs. The potential in modeling longer sequences can be seen from this study regardless of whether the particular analyses offered here are correct. We have not yet reached the limits of what we can learn using persistence, 1-back or N-back, as a tool for the investigation of sociolinguistic variation. By investigating quantitative patterns at the microtemporal level, we can learn more about what factors are at play in the production of sociolinguistic variation.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Abramowicz, L. (2007). "Sociolinguistics meets Exemplar Theory: frequency and recency effects in (ing)," in *University of Pennsylvania Working Papers in Linguistics*, Vol. 13 (Philadelphia, PA), 27–37.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., and Urbach, T. P. (1995). Syntactic priming: investigating the mental representation of language. *J. Psycholinguist. Res.* 24, 489–506. doi: 10.1007/BF02143163

Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behav. Res. Methods* 41, 977–90. doi: 10.3758/BRM.41.4.977

Estival, D. (1985). Syntactic priming of the passive in English. *Text* 5, 7–21. doi: 10.1515/text.1.1985.5.1-2.7

Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *J. Psycholinguist. Res.* 34, 365–399. doi: 10.1007/s10936-005-6139-3

Labov, W. (2001). "Chapter 5: The anatomy of style-shifting," in *Style and Sociolinguistic Variation*, eds P. Eckert and J. R. Rickford (Cambridge, UK: Cambridge University Press), 85–108.

Labov, W. and Rosenfelder, I. (2011). *The Philadelphia Neighborhood Corpus* of LING 560 studies, 1972-2010. With support of NSF contract 921643.

Poplack, S. (1980). Deletion and disambiguation in Puerto Rican Spanish. *Language* 56, 371–385. doi: 10.1353/lan.1980.0033

Poplack, S. (1984). Variable concord and sentential plural marking in Puerto Rican Spanish. *Hispanic Rev.* 52, 205–222. doi: 10.2307/473375

Poplack, S., and Tagliamonte, S. (1989). There's no tense like the present: verbal -s inflection in early Black English. *Lang. Variat. Change* 1, 47–84. doi: 10.1017/S0954394500000119

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). *FAVE Program Suite [Forced Alignment and Vowel Extraction]*. University of Pennsylvania.

Sankoff, D., and Laberge, S. (1978). "Chapter 8: Statistical dependence among successive occurrences of a variable in discourse," in *Linguistic Variation: Models and Methods*, ed D. Sankoff (Cambridge, MA: Academic Press), 119–126.

Scherre, M. (2001). Phrase-level parallelism effect on noun phrase number agreement. *Lang. Variat. Change* 13, 91–107. doi: 10.1017/S0954394501131042

Scherre, M., and Naro, A. (1991). Marking in discourse: "Birds of a feather". *Lang. Variat. Change* 3, 23–32. doi: 10.1017/S0954394500000430

Scherre, M., and Naro, A. (1992). The serial effect on internal and external variables. *Lang. Variat. Change* 4, 1–13.

Szmrecsanyi, B. (2006). *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin: Mouton de Gruyter.

Tamminga, M. (2014). *Persistence in the production of linguistic variation* (Ph.D. thesis). University of Pennsylvania, Philadelphia, PA.

Tamminga, M. (2016). Persistence in phonological and morphological variation. *Lang. Variat. Change* 28, 335–356. doi: 10.1017/S0954394516000119

Tamminga, M., Ahern, C., and Ecay, A. (2016). Generalized Additive Mixed Models for intraspeaker variation. *Linguist. Vanguard* 2, 33–41 doi: 10.1515/lingvan-2016-0030

Weiner, E. J., and Labov, W. (1983). Constraints on the agentless passive. *J. Linguist.* 19, 29–58.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 12 July 2019
doi: 10.3389/frai.2019.00011

Mapping Lexical Dialect Variation in British English Using Twitter

Jack Grieve[1]*, *Chris Montgomery*[2], *Andrea Nini*[3], *Akira Murakami*[1] *and Diansheng Guo*[4]

[1] Department of English Language and Linguistics, University of Birmingham, Birmingham, United Kingdom, [2] School of English, University of Sheffield, Sheffield, United Kingdom, [3] Department of Linguistics and English Language, University of Manchester, Manchester, United Kingdom, [4] Department of Geography, University of South Carolina, Columbia, SC, United States

There is a growing trend in regional dialectology to analyse large corpora of social media data, but it is unclear if the results of these studies can be generalized to language as a whole. To assess the generalizability of Twitter dialect maps, this paper presents the first systematic comparison of regional lexical variation in Twitter corpora and traditional survey data. We compare the regional patterns found in 139 lexical dialect maps based on a 1.8 billion word corpus of geolocated UK Twitter data and the BBC Voices dialect survey. A spatial analysis of these 139 map pairs finds a broad alignment between these two data sources, offering evidence that both approaches to data collection allow for the same basic underlying regional patterns to be identified. We argue that these results license the use of Twitter corpora for general inquiries into regional lexical variation and change.

Keywords: dialectology, social media, Twitter, British English, big data, lexical variation, spatial analysis, sociolinguistics

OPEN ACCESS

Edited by:
Jonathan Dunn,
University of Canterbury, New Zealand

Reviewed by:
Benedikt Szmrecsanyi,
KU Leuven, Belgium
Gabriel Doyle,
San Diego State University,
United States
John Nerbonne,
University of Freiburg, Germany

****Correspondence:***
Jack Grieve
j.grieve@bham.ac.uk

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *02 April 2019*
Accepted: *12 June 2019*
Published: *12 July 2019*

Citation:
Grieve J, Montgomery C, Nini A,
Murakami A and Guo D (2019)
Mapping Lexical Dialect Variation in
British English Using Twitter.
Front. Artif. Intell. 2:11.
doi: 10.3389/frai.2019.00011

INTRODUCTION

Regional dialectology has traditionally been based on data elicited through surveys and interviews, but in recent years there has been growing interest in mapping linguistic variation through the analysis of very large corpora of natural language collected online. Such corpus-based approaches to the study of language variation and change are becoming increasingly common across sociolinguistics (Nguyen et al., 2016), but have been adopted most enthusiastically in dialectology, where traditional forms of data collection are so onerous. Dialect surveys typically require fieldworkers to interview many informants from across a region and are thus some of the most expensive and complex endeavors in linguistics. As a result, there have only been a handful of surveys completed in the UK and the US in over a century of research. These studies have been immensely informative and influential, shaping our understanding of the mechanisms of language variation and change and giving rise to the modern field of sociolinguistics, but they have not allowed regional dialect variation to be fully understood, especially above the levels of phonetics and phonology. As was recently lamented in the popular press (Sheidlower, 2018), this shift from dialectology as a social science to a data science has led to a less personal form of scholarship, but it has nevertheless reinvigorated the field, democratizing dialectology by allowing anyone to analyse regional linguistic variation on a large scale.

The main challenge associated with corpus-based dialectology is sampling natural language in sufficient quantities from across a region of interest to permit meaningful analyses to be conducted. The rise of corpus-based dialectology has only become possible with the rise of computer-mediated communication, which deposits massive amounts of regionalized language data online every day. Aside from early studies based on corpora of letters to the editor downloaded from newspaper websites (e.g., Grieve, 2009), this research has been almost entirely based on Twitter, which facilitates the collection of large amounts of geolocated data. Research on regional lexical variation on American Twitter has been especially active (e.g., Eisenstein et al., 2012, 2014; Cook et al., 2014; Doyle, 2014; Jones, 2015; Huang et al., 2016; Kulkarni et al., 2016; Grieve et al., 2018). For example, Huang et al. (2016) found that regional dialect patterns on American Twitter largely align with traditional dialect regions, based on an analysis of lexical alternations, while Grieve et al. (2018) identified five main regional patterns of lexical innovation through an analysis of the relative frequencies of emerging words. Twitter has also been used to study more specific varieties of American English. For example, Jones (2015) analyzed regional variation in African American Twitter, finding that African American dialect regions reflect the pathways taken by African Americans as they migrated north during the Great Migration. There has been considerably less Twitter-based dialectology for British English. Most notably, Bailey (2015, 2016) compiled a corpus of UK Twitter and mapped a selection of lexical and phonetic variables, while Shoemark et al. (2017) looked at a Twitter corpus to see if users were more likely to use Scottish forms when tweeting on Scottish topics. In addition, Durham (2016) used a corpus of Welsh English Twitter to examine attitudes toward accents in Wales, and Willis et al. (2018) have begun to map grammatical variation in the UK.

Research in corpus-based dialectology has grown dramatically in recent years, but there are still a number of basic questions that have yet to be fully addressed. Perhaps the most important of these is whether the maps of individual features generated through the analysis of Twitter corpora correspond to the maps generated through the analysis of traditional survey data. Some studies have begun to investigate this issue. For example, Cook et al. (2014) found that lexical Twitter maps often match the maps in the *Dictionary of American Regional English* and *Urban Dictionary* (see also Rahimi et al., 2017), while Doyle (2014) found that Twitter maps are similar to the maps from the *Atlas of North American English* and the *Harvard Dialect Survey*. Similarly, Bailey (2015, 2016) found a general alignment for a selection of features for British English. While these studies have shown that Twitter maps can align with traditional dialect maps, the comparisons have been limited—based on some combination of a small number of hand selected forms, restricted comparison data (e.g., dictionary entries), small or problematically sampled Twitter corpora (e.g., compiled by searching for individual words), and informal approaches to map comparison.

A feature-by-feature comparison of Twitter maps and survey maps is needed because it is unclear to what extent Twitter maps reflect general patterns of regional linguistic variation. The careful analysis of a large and representative Twitter corpus is sufficient to map regional patterns on Twitter, but it is also important to know if such maps generalize past this variety, as this would license the use of Twitter data for general investigations of regional linguistic variation and change, as well as for a wide range of applications. The primary goal of this study is therefore to compare lexical dialect maps based on Twitter corpora and survey data so as to assess the degree to which these two approaches to data collection yield comparable results. We do not assume that the results of surveys generalize; rather, we believe that alignment between these two very different sources of dialect data would be strong evidence that both approaches to data collection allow for more general patterns of regional dialect variation to be mapped. A secondary goal of this study is to test how consistent dialect patterns are across different communicative contexts. Corpus-based dialectology has shown that regional variation pervades language, even in the written standard (Grieve, 2016), but we do not know how stable regional variation is on the level of individual linguistic features. To address these gaps in our understanding of regional linguistic variation, this paper presents the first systematic comparison of lexical dialect maps based on surveys and Twitter corpora. Specifically, we report the results of a spatial comparison of the maps for 139 lexical variants based on a multi-billion-word corpus of geocoded British Twitter data and the BBC Voices dialect survey.

BRITISH DIALECTOLOGY

Interest in regional dialect variation in Great Britain is longstanding, with the earliest recorded comments on accent dating back to the fifteenth and sixteenth centuries (Trevisa, 1495). The study of regional variation in lexis grew in popularity during the late eighteenth and early nineteenth centuries, with dialect glossaries being compiled across the country, especially in Yorkshire and the North, in order to preserve local lexis, which was assumed to be going extinct. Most notably, Wright's (1898) *English Dialect Dictionary*, which drew on many of these glossaries, detailed lexical variation across the British Isles, especially England. The earliest systematic studies of accents in England also began around this time (see Maguire, 2012).

It was not until the *Survey of English Dialects* (SED) (Orton, 1962), however, that a full survey of dialect variation across England was attempted. Data was collected between 1950 and 1961 in 313 primarily rural locations using a 1,322 question survey, which included 730 lexical questions. Respondents, typically older males who had lived most of their lives in that location, were interviewed face-to-face by a fieldworker. The rest of the UK was covered separately. Scotland and Northern Ireland, along with the far north of England, were mapped by *The Linguistic Survey of Scotland*, which began collecting data in 1952 through a postal questionnaire (Mather et al., 1975). This survey also mapped regional variation in Scottish Gaelic (O'Dochartaigh, 1994). Finally, both Welsh (Jones et al., 2000) and English (e.g., Parry, 1999) in Wales were mapped in the late twentieth century.

With the rise of sociolinguistics in the 1960s and 1970s, work on language variation and change in the UK shifted focus from regional patterns to social patterns, generally based on interviews with informants from a range of social backgrounds from a single location. Interest in regional dialects, however, began to re-emerge recently. Llamas (1999) developed the Survey of Regional English (SuRE) to collect data from across levels of linguistic analysis. A national survey was never conducted, but the SuRE method was adopted for research in individual locations, including by Llamas (2007) in Middlesbrough, Asprey (2007) in the Black Country, and Burbano-Elizondo (2008) in Sunderland. In addition, the lexical component of the SuRE system was adapted for a national survey conducted as part of the BBC Voices project (Elmes, 2013). BBC Voices was designed to provide a snapshot of modern language use in the UK and employed various methods for data collection, including group interviews (Robinson et al., 2013), an attitudinal questionnaire (Bishop et al., 2005), and a web-based survey to collect lexical data based on SuRE. This lexical data, discussed below, is the basis for the present study. It has previously been subjected to statistical analysis (Wieling et al., 2014), which found evidence for four dialect regions (Southern England, Northern England, Scotland, and Northeast Scotland) based on a multivariate analysis of the maps for the top 10 variants of each of the 38 alternations. In addition to the BBC Voices survey, three other UK dialect surveys have recently come online. In 2007, Bert Vaux initiated the Cambridge online survey of World Englishes, which collects data on 31 alternations of various types from across the world, including the UK. MacKenzie et al. (2015) collected data on 31 alternations of various types from across the UK, with the help of undergraduate Linguistics and English Language students at the University of Manchester. Finally Leemann et al. (2018) used a mobile phone app to collect data on 26 alternations, primarily related to pronunciation, from over 47,000 speakers from over 4,900 localities from across the UK.

There is also a long history of corpus-based research in British dialectology. Most research on Old and Middle British dialects is essentially corpus-based, as it relies on samples of historical writing (e.g., Brook, 1963), but more specifically dialect corpora were compiled to map regional patterns in contemporary British English in the 1970s and 1980s. The first was the 1 million word *Helsinki Corpus of British English Dialects* (Ihalainen et al., 1987), designed as a grammatical supplement to the SED. Informants were recorded in their home and encouraged to talk about any subject they pleased to elicit naturalistic speech. The second was the 2.5 million word *Freiburg Corpus of English Dialects*, which contains transcriptions of interviews with older informants telling their life stories to fieldworkers (see Anderwald, 2009; Szmrecsanyi, 2013). Because these datasets consist of transcriptions of interviews elicited from a small number of informants, they fall in between traditional dialect surveys and the large natural language corpora that are the focus of this study.

Despite this long tradition of research, relatively little is known about regional linguistic variation in contemporary British English, especially compared to American English and especially in regard to lexical and grammatical variation. In large part this is because so few researchers have yet to take advantage of the immense social media corpora that can now be compiled and whose popularity is driving dialectology around the world. In addition to comparing lexical variation in corpora and surveys, a secondary goal of this study is therefore to encourage the adoption of computational approaches in British dialectology.

MATERIALS AND METHODS

BBC Voices Dataset

The regional dialect survey data we used for this study was drawn from the BBC Voices project (Upton, 2013)[1]. We chose this dataset, which was collected online between 2004 and 2007, not only because it is easily accessible, but because it is the most recent lexical dialect survey of British English and because it focuses on everyday concepts, whereas older surveys tended to focus on archaic words and rural concepts, which are rarely discussed on Twitter.

The BBC Voices survey collected ∼734,000 responses from ∼84,000 informants to 38 open-ended questions, each designed to elicit the variants of a lexical alternation. The criteria for the selection of these 38 questions is unclear. Some (e.g., what word do you use for *running water smaller than a stream*) had been included in previous surveys, whereas others (e.g., *young person in cheap trendy clothes and jewelery*) were seemingly intended to elicit emerging forms (i.e., *chav*). In addition, two questions (*male partner*, *female partner*) are associated with variants that are not generally interchangeable (e.g., *boyfriend/husband*, *girlfriend/wife*); we therefore excluded these questions from our final analysis. All informants did not respond to all questions. The most responses were provided for *drunk* (29,275) and the fewest for *to play (a game)* (9,897). Across all responses, 1,146 variants were provided, with the most for *drunk* (104) and the fewest for *mother* (10). For example, of the 18 variants supplied in the 11,272 responses to the *left-handed* question, *cack-handed* (4,101) and *left* (3,987) are most common, together accounting for 72% of responses.

The large number of variants associated with each alternation is problematic because if we considered the complete set, our comparison would be dominated by very uncommon forms, which cannot be mapped accurately. Consequently, we only considered the most common variants of each alternation. In doing so, however, we violated the *principle of accountability*, which requires all variants to be taken into consideration (Labov, 1972). Fortunately, this frequency distribution ensures that excluding less common variants, which contribute so few tokens, will have almost no effect on the proportions of the more common variants. We therefore only retained variants that were provided by at least 5% of respondents. We tested other cut-offs, but higher thresholds (e.g., 10%) resulted in variants with clear regional patterns being excluded, whereas lower thresholds (e.g., 1%) resulted in variants that are too infrequent to show patterns being included.

Not only is each alternation associated with multiple variants, but each variant is associated with multiple distinct orthographic forms. These are the specific answers provided by informants that were judged by the BBC Voices team to be closely related to that variant, including inflections, non-standard spellings, and multiword units. Across all responses, 45,573 distinct forms were provided (ignoring capitalization), with the most for *unattractive*

[1] We downloaded the BBC Voices survey dataset, specifically the 'RADAR 1' component of the dataset, from the project website, which is available at http://routledgetextbooks.com/textbooks/_author/upton-9780415694438

(2,300) and the fewest for *a long seat* (285). For example, of the 4,101 *cack-handed* responses to the *left-handed* question, informants provided 142 distinct orthographic forms, including "cack handed" (1,833) and "cack-handed" (1,026), which account for 70% of all responses, with the 18 forms provided by at least 10 informants accounting for 95% of responses. Alternatively, there are 86 forms provided by one informant, including "kerhandit" and "cack handedEnter Word," the latter form clearly representing a data entry error.

The large number of forms associated with each variant is also problematic, especially because many of the most uncommon forms are of unclear status. This includes not only data entry errors, but forms that are almost never used with the target meaning, such as "china" for *mate*, which comes from "china plate" in Cockney rhyming slang. Fortunately, the frequency distribution also allowed us to exclude less frequent forms from our analysis without affecting the regional patterns of more frequent variants. For each variant we only included forms that were returned by at least 50 informants.

At the end of this process, our final feature set includes 36 alternations (e.g., left-handed), associated with 139 variants (e.g., *cack-handed*, *left*, *cag-handed*), which in turn are associated with 291 distinct orthographic forms (e.g., *cack handed, cack-handed*, etc.). The complete set of alternations and variants is presented in **Table 1**. The complete set of forms are included in the **Supplementary Materials**. The number of variants per alternation ranges from 2 to 7, most with 4 variants; the number of forms per variant ranges from 1 to 12, most with 2 forms. Notably, there are 291 forms in our dataset, but only 288 unique forms, because 3 are linked to the variants of multiple alternations: "chuck" is associated with the *throw* and *heavy rain* alternations, "hot" with the *hot weather* and *attractive* alternations, and "pissed" with the *annoyed* and *drunk* alternations. This situation is problematic and points to a larger issue with polysemy (and homophony) in our feature set, which we return to later in this paper, but crucially because the proportional use of each variant is calculated relative to the frequency of the other variants of that alternation, the maps for these overlapping variants are distinct.

After selecting these 139 variants, we extracted the regional data for each from the BBC Voices dataset, which provides the percentage of informants in 124 UK postal code areas who supplied each variant. For example, the *cack-handed* variant accounted for 4,101 out of the 11,272 responses for the *left-handed* alternation (36%), with a minimum of 0% of informants using this form in the Shetlands and a maximum of 100% of informants in Jersey. Notably, these two extreme postal code areas have the fewest respondents, leading to generally less reliable measurements for these areas. Most areas, however, are associated with far more informants and thus exhibit much more variability. For example, 96% of postal code areas are characterized by between 10 and 70% usage of this particular variant. There are also a very small number of missing data points in our BBC Voices dataset (48 out of 17,236 values), which occur in cases where no responses were provided by any informants in that postal code area for that question. Because this is a negligible amount of missing data and because it is distributed across many variants, we simply assigned the mean value for that variant across all locations to those locations. In addition, because the BBC Voices dataset provides percentages calculated based on the complete set of variants, whereas we are looking at only the most common variants, we recalculated the percentage for each variant in each postal code area based only on the variants selected for analysis. For example, in the Birmingham area, the overall percentages for *cack-handed* (32.3%), *left* (23.8%), and *cag-handed* (32%), which cumulatively account for 88.1% of responses, were recalculated as 36.7, 27, and 36.3%, respectively, which sum to 100%.

Finally, we mapped each of the variants in this dataset. For example, the maps for the alternation between *sofa/couch/settee* is presented in the first column of **Figure 1**, where each map plots the percentage of one variant across the 124 postal code areas in the BBC Voices dataset. In this case, a clear regional pattern can be seen within and across variants, with *sofa* being relatively more common in the South, *couch* in Scotland, and *settee* in the Midlands and the North of England. The complete set of maps are presented in the **Supplementary Materials**.

UK Twitter Dialect Corpus

The regional dialect corpus used for this study consists of a large collection of geolocated Twitter data from the UK that we downloaded between 2014-01-01 and 2014-12-31 using the Twitter API. This data was collected as part of a larger project that has explored lexical variation on Twitter (see also Huang et al., 2016; Grieve et al., 2017, 2018; Nini et al., 2017). In total, this corpus contains 1.8 billion words, consisting of 180 million Tweets, posted by 1.9 million unique accounts. The median number of Tweets per account is 10. The corpus contains data for 360 days, with data for 5 days missing due to technical issues. To analyse regional variation in the corpus, we formed regional sub-corpora by grouping all individual Tweets by postal code regions based on the provided longitude and latitude. Postal code regions were used to facilitate comparison with the BBC Voices data. Overall, the corpus contains 124 postal code regions, with on average 1.5 million Tweets per region, with the number of Tweets varying from between 5.5 million Tweets in Manchester to 54,000 Tweets in the Outer Hebrides, reflecting variation in population; London is not the largest region because it is subdivided into smaller areas.

Notably, we do not filter our corpus in any way, for example by excluding re-Tweets or spam or Tweets from prolific posters or bots. Tweets from one user may also appear in different regional sub-corpora if the user was in different postal code regions when those posts were made. The Twitter corpus analyzed in this study is an unbiased sample of geolocated Tweets, similar to what a user would see if they browsed Tweets from a region at random. We believe that modifying the corpus to make it more likely to show regional patterns is a highly subjective process that necessarily results in a less representative corpus. By including all Tweets from a given region in our corpus, we have taken a more conservative choice, allowing us to assess the base level of alignment between Twitter data and traditional dialect surveys. Removing Tweets from the corpus may lead to the identification of stronger regional patterns or better alignment

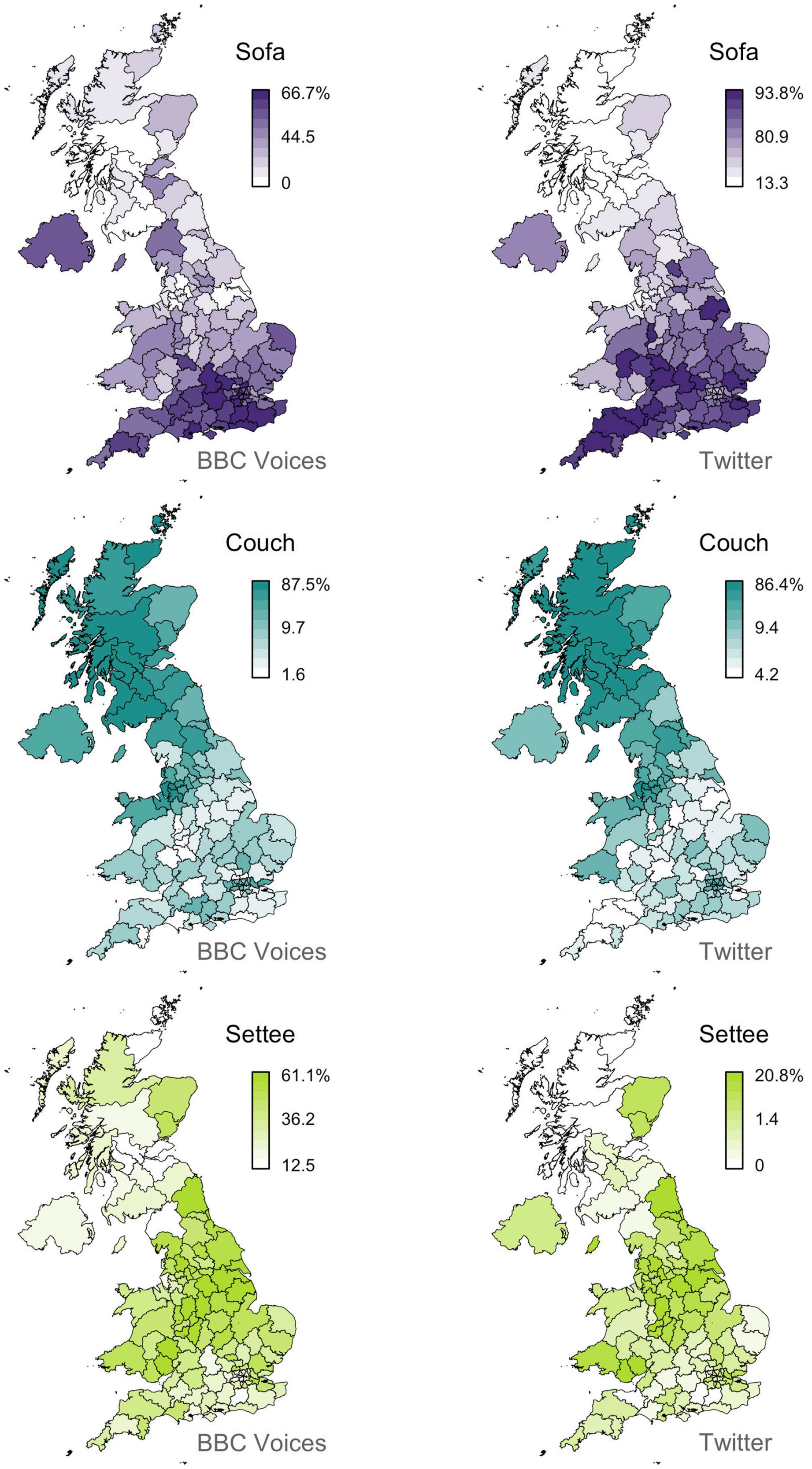


FIGURE 1 | *Sofa/couch/settee* alternation.

TABLE 1 | Feature set.

	Alternation	Total	Variants
1	Hot	6	*Boiling, roasting, hot, baked, sweltered, sweating*
2	Cold	4	*Freezing, chilly, nippy, cold*
3	Tired	2	*Knackered, shattered*
4	Unwell	3	*Sick, poorly, ill*
5	Pleased	3	*Chuffed, happy, made up*
6	Annoyed	2	*Pissed off, angry*
7	To play a game	2	*Play, lake*
8	To play truant	5	*Skive, bunk, wag, play hookey, skip*
9	Throw	2	*Chuck, lob*
10	Hit hard	5	*Whack, smack, thump, wallop, belt*
11	Sleep	5	*Kip, sleep, snooze, nap, doze*
12	Drunk	2	*Pissed, wasted*
13	Pregnant	4	*Up the duff, pregnant, bun in the oven, expecting*
14	Left-handed	3	*Cack-handed, left, cag-handed*
15	Lacking money	4	*Skint, broke, poor, brassic*
16	Rich	5	*Loaded, minted, well off, rolling in it, rich*
17	Insane	5	*Mad, nuts, crazy, mental, bonkers*
18	Attractive	4	*Fit, gorgeous, pretty, hot*
19	Unattractive	2	*Ugly, minger*
20	Moody	4	*Mardy, grumpy, stroppy, moody*
21	Baby	7	*Baby, bairn, sprog, babby, kid, wean, little one*
22	Mother	5	*Mum, mam, mummy, ma, mom*
23	Grandmother	3	*Nanny, granny, grandma*
24	Grandfather	4	*Grandad, grandpa, grampa, pop*
25	Friend	4	*Mate, pal, friend, buddy*
26	Young person in cheap trendy clothes and jewelery	4	*Chav, townie, scally, ned*
27	Clothes	5	*Clothes, gear, clobber, togs, kit*
28	Trousers	5	*Trousers, pants, keks, jeans, trews*
29	Child's soft shoes worn for PE	4	*Plimsolls, pumps, daps, trainers*
30	Main room of house (with TV)	4	*Living room, lounge, sitting room, front room*
31	Long soft seat in the main room	3	*Sofa, settee, couch*
32	Toilet	4	*Loo, bog, toilet, lavatory*
33	Narrow walkway alongside buildings	4	*Alley, ginnel, pavement, path*
34	To rain lightly	3	*Drizzle, spit, shower*
35	To rain heavily	4	*Pour, piss, chuck, bucket*
36	Running water smaller than a river	4	*Stream, brook, burn beck*
		139	

with dialect survey maps, but this can only be tested once a baseline is established.

Next, we measured the frequency of each of the 139 lexical variants in our BBC Voices dataset across our 124 postal code area sub-corpora. We then summed the counts for all forms associated with each variant in each postal code area and computed a percentage for each variant for each alternation in each postal code area by dividing the frequency of that variant by the frequency of all variants of that alternation in that postal code area. In this way, we created a regional linguistic dataset based on our Twitter corpus that matches our BBC Voices dataset, consisting of percentages for all 139 variants, grouped into 36 alternations, measured across the 124 postal code areas, where the percentages for the variants for each alternation sum to 100% in each postal code area. We also mapped the percentages of all 139 variants across the 124 postal code areas. For example, the Twitter maps for the alternation between *sofa*/*couch*/*settee* are presented in the second column of **Figure 1**. The complete set of maps are presented in the **Supplementary Materials**.

Crucially, we counted all tokens of the variants in our corpus, making no attempt to disambiguate between word senses. For example, the variant *spit* in the alternation between *drizzle*/*spit* is used more often in the corpus to refer to the physical action as opposed to light rain, but we counted all tokens of *spit* regardless of the meaning it expressed. This is the simplest and most common approach in Twitter-based dialectology, although it is clearly problematic. Automatic word sense disambiguation systems are not commonly used in corpus-based dialectology because they are difficult to apply at scale and are fairly inaccurate, especially when working with uncommon dialect forms in highly informal data. We return to the issue of polysemy later in this paper, when we consider how variation in meaning affects the overall alignment between the two sets of maps and how much alignment can be improved through the application of techniques for word sense disambiguation.

Finally, it is important to acknowledge that Twitter corpora do not represent language in its entirety. Twitter corpora only represent Twitter, which is a very specific form of public, written, computer-mediated communication. The unique constellation of situational properties that define Twitter affect its form and differentiate it from other varieties of languages, as does the demographic background of Twitter users, who in the UK are more likely to be young, male, and well-educated compared to the general population (Longley et al., 2015; Mellon and Prosser, 2017). These are the social and situational patterns that define Twitter and they should be reflected in any corpus that attempts to represent this variety of language. The goal of this study is to evaluate the degree to which general patterns of regional variation persist in Twitter corpora despite its unique characteristics.

Lee's *L*

To systematically assess the similarity of the Twitter maps and the survey maps we measured the degree of alignment between each pair of maps. There is, however, no standard method for bivariate map comparison in dialectology. Other than visually comparing dialect maps (e.g., Grieve et al., 2013), the simplest approach is to correlate the two maps by calculating a correlation coefficient (e.g., Pearson's *r*), essentially comparing the values of the two maps at every pair of locations. This was the approach taken in Grieve (2013), for example, where Pearson correlation coefficients were calculated to compare a small number of maps representing general regional patterns of grammatical and phonetic variation. This is also the general approach underlying many dialect studies that have used methods like factor analysis (e.g., Nerbonne, 2006) and principal components analysis (e.g., Shackleton, 2007) to identify common regional patterns in large sets of dialect maps based on correlation (or covariance) matrices. Although correlating dialect maps generally appears to yield consistent and meaningful results, this process ignores the spatial distribution of the values of each variable. Consequently, the similarity between two dialect maps can be estimated incorrectly and significance testing is unreliable, as it is based on the assumption that the values of a variable are independent across locations (see Lee, 2001).

Alternatively, methods in spatial analysis have been designed specifically for inferential bivariate map comparison (Wartenberg, 1985; Lee, 2001). Most notably, Lee (2001) proposed a spatial correlation coefficient (*L*) that measures the association between two geographically referenced variables, taking into account their spatial distribution. Lee's *L* is essentially a combination of Pearson's *r*, the standard bivariate measure of association, and Moran's *I*, the standard univariate measure of global spatial autocorrelation (see Grieve, 2018). On the one hand, Pearson's *r* correlates the values of two variables (x and y) by comparing the values of the variables at each pair of observations (i.e., locations) and can be expressed as

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

On the other hand, Moran's I compares the values of a single variable (x) across all pairs of locations, with the spatial distribution of the variable used to define a spatial weights matrix (w), which specifies the weight assigned to the comparison of each pair of locations (i, j). For example, a spatial weights matrix is often set at 1 for neighboring locations and 0 for all other pairs of locations. When row standardized, Moran's I can be expressed as

$$I = \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}$$

Combining these two measures, Lee defined his bivariate measure of spatial association L as

$$L = \frac{\sum_i \left(\left(\sum_j w_{ij}(x_j - \overline{x}) \right) \left(\sum_j w_{ij}(y_j - \overline{y}) \right) \right)}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

so that every pair of locations is compared within and across the two variables, taking into consideration the geographical distribution of the values. Like Pearson's r, Lee's L can range from -1 to $+1$, where stronger positive values indicate stronger matches. Lee's L is independent of scale, which is important as our maps can differ in terms of scale. In addition, pseudo-significance testing can be conducted for Lee's L through a randomization procedure, in much the same way as Moran's I. Lee's L is recalculated for a large number of random rearrangements of the locations over which the variable was measured. The set of values that results from this process represents the null distribution of Lee's L. The observed value of Lee's L is then compared to this null distribution to generate a pseudo p-value.

Finally, to calculate Lee's L, a spatial weights matrix must be defined. For this study, we used a nearest neighbor spatial weights matrix, where every location is compared to its nearest n neighbors, including itself, with each of these n neighbors assigned a weight of 1 and all other locations assigned a weight of 0. Following Grieve (2017), who suggests setting n at ∼10% of the total locations, our main analysis is based on 10 nearest neighbors, calculated using postal code area centroids, but we also ran the analysis based on 2, 5, and 20 nearest neighbors, so as to judge how sensitive our results are to this setting.

RESULTS

Map Comparison

We correlated all 139 pairs of Twitter and BBC Voices dialect maps using Lee's L, based on a 10 nearest neighbor spatial weights matrix. The 139 L values range from -0.28 to $+0.74$, with a median of +0.14, indicating a tendency for the maps to align. Overall, 109 of the 139 comparisons (78%) exhibit positive correlation coefficients, and 93 of these pairs (67%) exhibit significant correlations at the $p < 0.05$ level[2]. Computing Lee's L using 2, 5, and 20 nearest neighbors produced similar results, with all analyses finding that 78–80% of the map pairs exhibit positive correlations, and with the Lee's L values across all 139 pairs of maps exhibiting strong correlations ($r > 0.89$), indicating the choice of spatial weights matrix does not have a large effect on our results. We also computed Pearson correlation coefficients for all 139 pairs of maps, which yielded similar results (median $r = 0.22$, 82% of comparisons with positive correlations). Finally, there is a strong correlation between Pearson's r and Lee's L ($r = 0.90$), indicating that Lee's spatial adjustment does not have a large effect on our results.

These results demonstrate that the regional patterns in the BBC Voices survey data and our Twitter corpus are broadly comparable. It is unclear, however, just how similar these maps really are. Significant alignment, at any level, is not a guarantee of meaningful alignment. Furthermore, given standard rules of thumb for Pearson's r, a median Lee's L of 0.14 does not seem especially strong. We do not know, however, how exactly to interpret Lee's L within the context of this study. Ultimately, the question we are interested in answering is whether two sets of maps under comparison tend to align in a meaningful way for dialectologists. It is therefore crucial that we compare the two sets of maps visually to assess the degree of alignment, especially those map pairs that show seemingly low-to-middling correlations. In other words, we believe it is important to calibrate our interpretation of Lee's L for dialectological inquiry, rather than simply noting that a certain percentage of map pairs show a significant or substantial spatial correlation.

For example, we believe it is clear that the maps for *sofa*, *couch* and *settee* presented in **Figure 1** broadly align. Lee's correlation coefficients here range between $L = 0.63$ for *couch*, which is the eighth best match in our dataset, to $L = 0.27$ for *settee*, which is the 40th best match. Crucially, the result for *settee* suggests that what appears to be low-to-middling values for Lee's L might represent very meaningful alignments in the context of dialectology. To investigate this issue further, we examined how the visual similarity between the 139 pairs of maps degrades as Lee's L falls.

In **Figure 2**, we present 8 pairs of maps with L values ranging from 0.74 to 0.03. We can clearly see that the alignment between the two sets of maps falls with Lee's L, as expected. For example, the maps for *granny* ($L = 0.74$) show very similar patterns, identifying Scotland, Northern Ireland, and the Southwest as hotspots for this variant. The other three pairs of maps with $L > 0.4$ also appear to be very good matches. Below this level, we still find clear broad alignment between the maps, including for *mate* ($L = 0.24$), which is more common in England especially in the Midlands, and *scally* ($L = 0.17$), which is more common in the North, especially around Liverpool. Only *bonkers* ($L = 0.04$) shows no obvious alignment, but the two maps both show relatively little spatial clustering in the first place, and even these maps are not obviously inconsistent with each other. In **Figure 3**, we present 8 pairs of maps with L values around 0.14—the median Lee's L across all 139 maps. Once again, we see broad alignment across the maps, although there is considerably more local variation than most of the pairs of maps presented in **Figure 2**. For example, *chuck* ($L = 0.15$) is identified as occurring primarily outside England in both maps, but the Twitter map is less definitive and also identifies a hotspot in the Southwest. *Sick* ($L = 0.13$) probably shows the worst overall match across these 8 examples: both maps show the form is relatively common in Northern Ireland and the Southeast, but only the BBC Voices map also identifies Scotland as a hotspot. Finally, in **Figure 4**, we present 8 pairs of maps with p values around 0.05, all of which are associated with L values of <0.1. There is at least partial alignment between all pairs of maps associated with $p < 0.05$. For example, both maps identify *grandpa* ($L = 0.06$, $p = 0.01$) as occurring relatively more often in Scotland and the Home Counties, although the status of Northern Ireland and Wales is inconsistent. Even the maps for *spit* ($L = 0.06$, $p = 0.06$) align to some degree, with both identifying hotspots around Liverpool.

Overall, we therefore find considerable alignment between the BBC Voices and the Twitter lexical dialect maps. The matches are far from perfect, but in our opinion a clear majority of the map pairs analyzed in this study show real correspondence, with the nations of the UK and the major regions of England being generally classified similarly in both sets of maps. The maps do not appear to be suitable in most cases for more fine-grained interpretations, except at higher levels of correlation, but given that these maps are defined at the level of postal code areas, which in most cases are fairly large regions, this seems like a reasonable degree of alignment, suggesting that these two approaches to data collection in dialectology allow for similar broad underlying patterns of regional lexical variation to be identified in British English.

Understanding Misalignments

Although the Twitter maps and the survey maps broadly correspond, the degree of alignment varies considerably across the 139 map pairs. To understand why some Twitter maps match the survey maps better than others, we considered how well alignment is predicted by three factors: the frequency of each variant in the Twitter corpus, the amount of spatial clustering in each Twitter map, and the likelihood of each variant occurring with the target meaning in the Twitter corpus. Knowing how these three characteristics of Twitter maps predict their alignment with survey maps not only offers guidance for improving the accuracy of Twitter maps, but it provides a basis for judging if new Twitter maps are likely to generalize, without comparison to survey maps, which are unavailable for most lexical alternations.

[2] We did not adjust the p-value for multiple comparisons because our goal is not to identify individual pairs of maps that show significant correlations. Rather, we are interested in reporting the proportion of the 139 map pairs that show a meaningful level of correlation in the context of dialectological inquiry, which is a much stricter test of robustness.

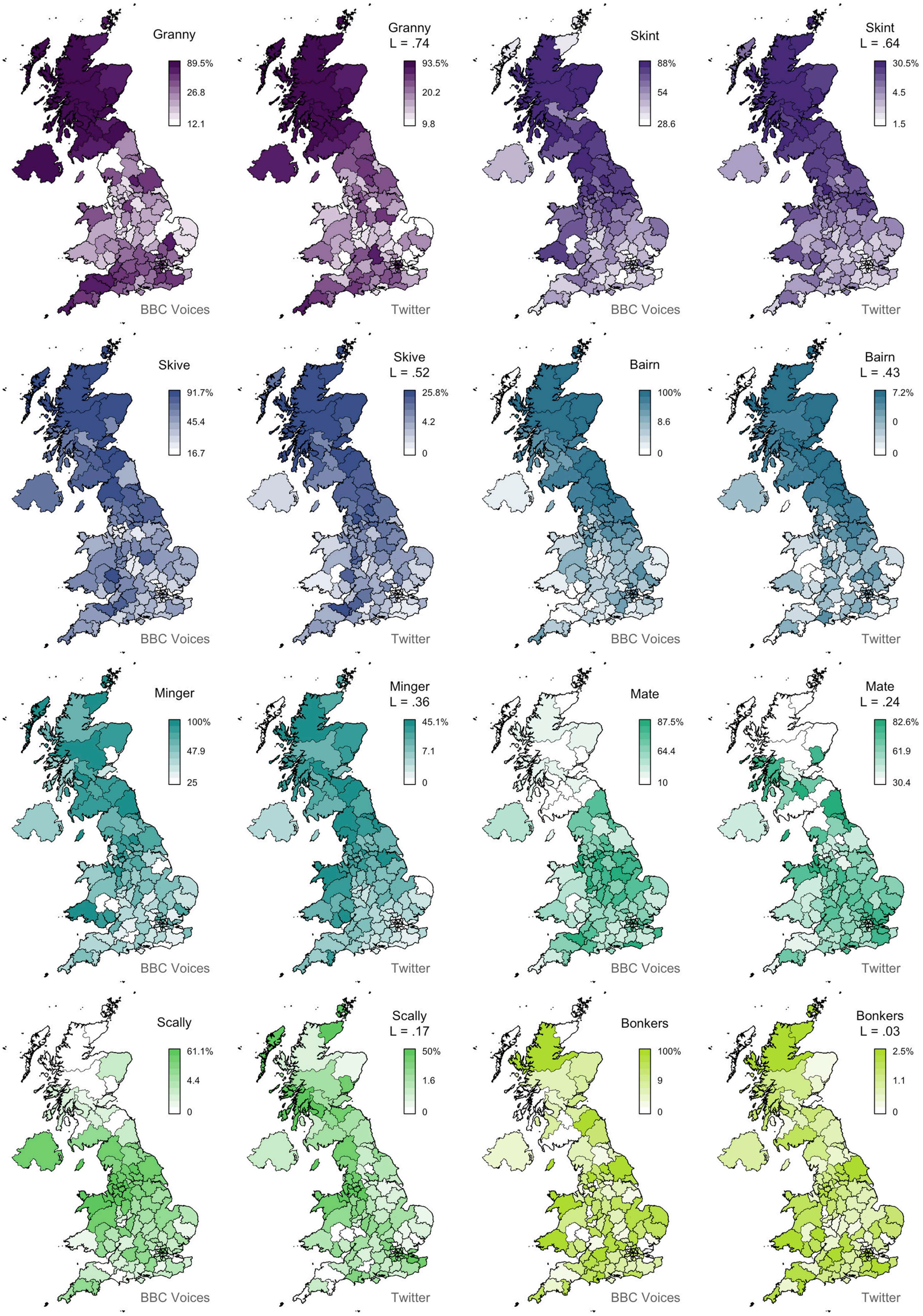


FIGURE 2 | Map comparisons (part 1).

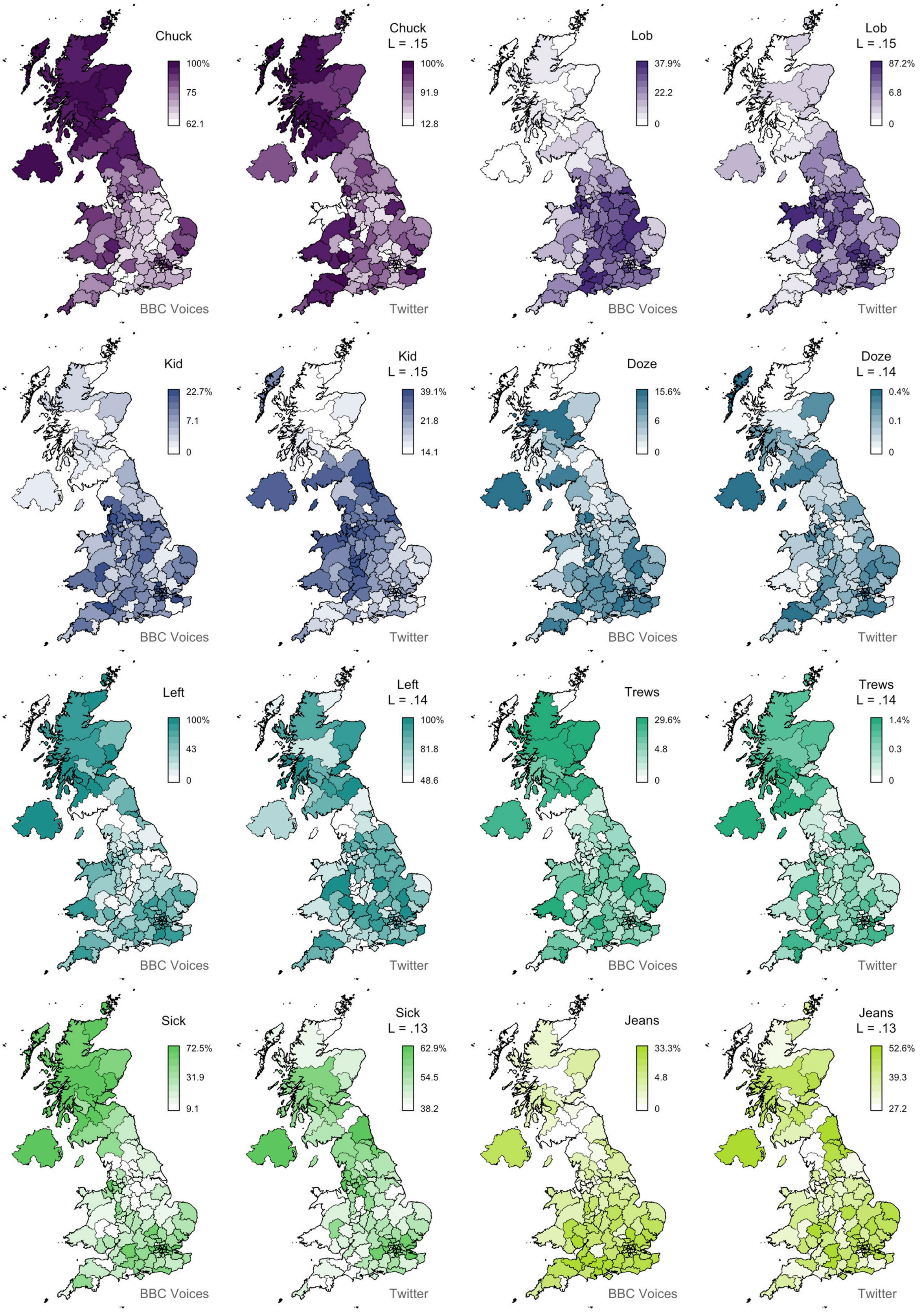


FIGURE 3 | Map comparisons (part 2).

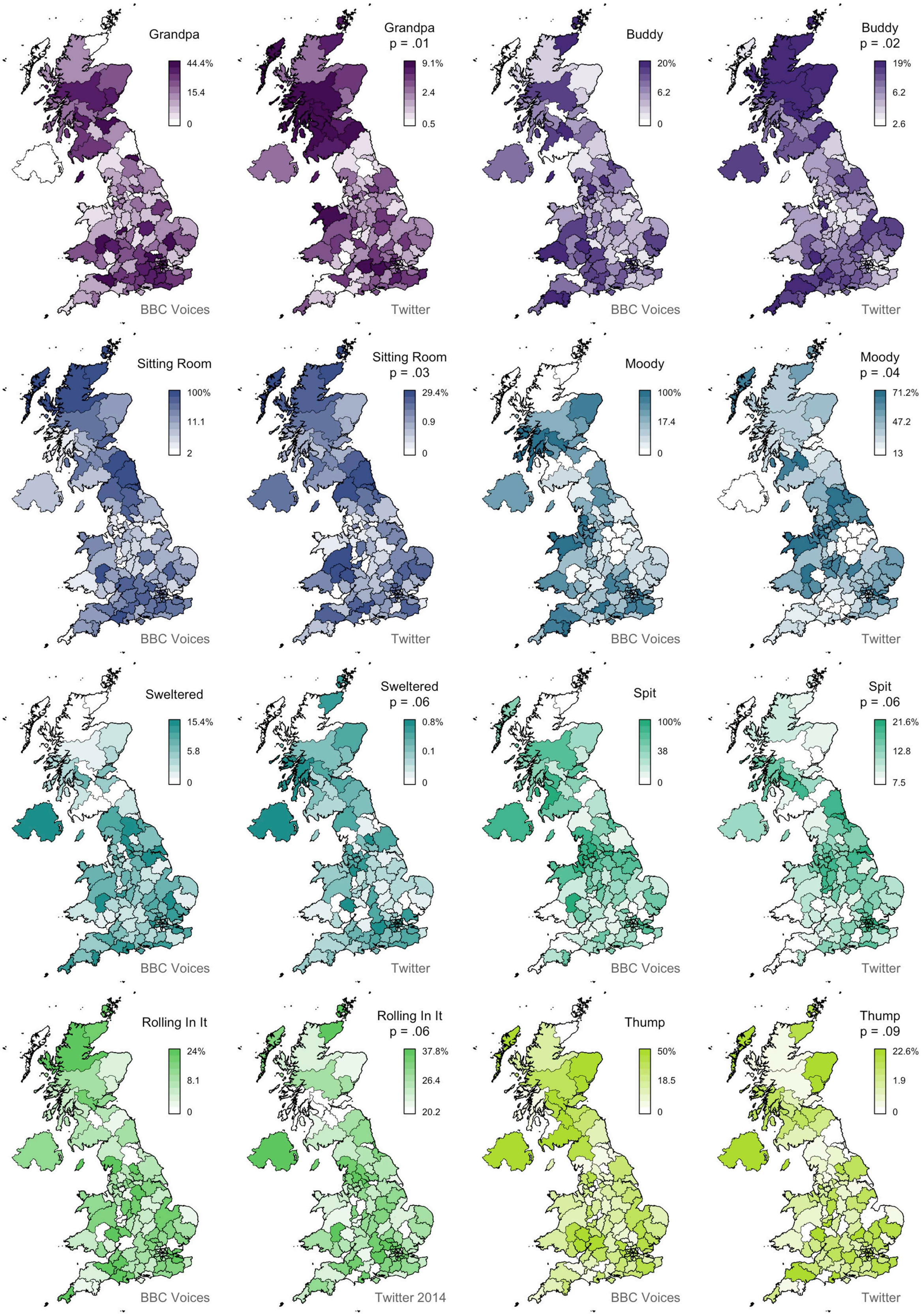


FIGURE 4 | Map comparisons (part 3).

TABLE 2 | Summary of the mixed-effects model fitted to Lee's *L*.

	Parameter	Estimate	SE	Standardized estimate
Fixed effects	Intercept	0.0412	0.0796	0.1648
	Moran's I	0.8172***	0.0845	0.1579
	Log-transformed frequency	−0.0250**	0.0075	−0.0532
	Target meaning ratio	0.0010*	0.0004	0.0357
Random effects	SD of random intercepts	0.1220		

**p* < 0.05, ***p* < 0.01, ****p* < 0.001, p-values calculated using Satterthwaite's approximation.*

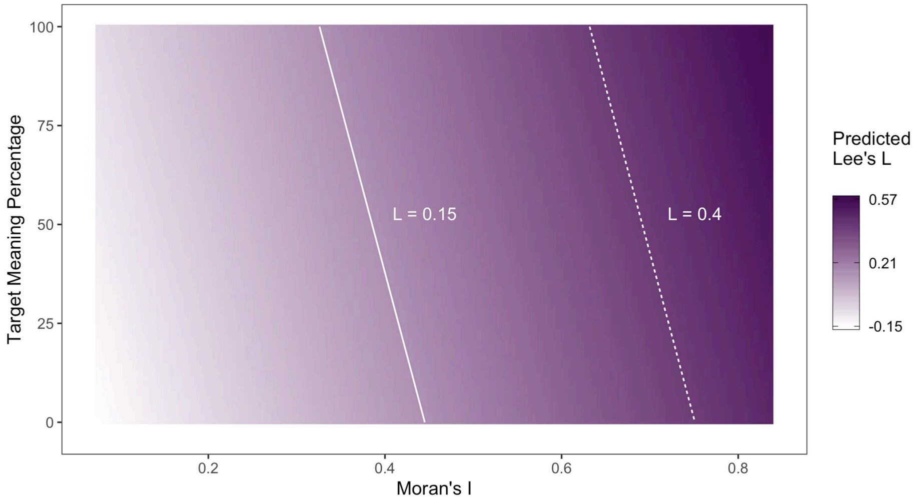


FIGURE 5 | Expected value of Lee's *L* as a function of Moran's I and target meaning ratio.

First, we included the frequency of each of the 139 variants in the complete Twitter corpus as a predictor in our model based on the assumption that measures of relatively frequency become better estimates of their true values as the number of tokens seen increases. Our intent was to assess how much misalignment can be explained by Twitter maps being based on too few observations. Second, we included the strength of the regional pattern exhibited by each of the 139 Twitter maps as a predictor in our model by computing the global spatial autocorrelation statistic Moran's *I* for each Twitter map using a 10 nearest neighbor spatial weights matrix. Our intent was to assess how much map misalignment can be explained by Twitter maps failing to exhibit clear regional patterns. It is important to acknowledge, however, that if the survey maps also fail to show regional patterns, misalignment should not be interpreted as evidence that the Twitter maps are inaccurate, as two random maps should not be expected to align. Furthermore, in general we expect these two measures to be correlated, as we know that Moran's *I* forms part of the foundation for Lee's *L*. Nevertheless, we wanted to assess how strong this relationship is, how much alignment increases with spatial clustering, and how much variation is left to be explained by other factors. Finally, we included an estimate of the percentage of tokens that were used with the target meaning in the corpus for each of the 139 variants as a predictor in our model by extracting 50 random concordance lines for each variant and coding them as target or non-target uses. Although polysemy is not an issue in surveys, where informants are asked to name concepts, variation in meaning should affect the accuracy of our Twitter maps, which were based on counts for all tokens of a variant regardless of their meaning. Our intent was to assess how much map misalignment is due to variation in the meaning of variants in the Twitter corpus.

We fit a linear mixed-effects regression model to Lee's *L*, measured across the 139 map pairs, with log-transformed frequency, Moran's *I*, and the percentage of target meaning as predictors, including alternation as a random intercept to account for the fact that the 139 variants are grouped into 36 alternations. Parameters were estimated using restricted maximum likelihood. Although Lee's *L* can range from −1 to +1, we used a linear model because the observed values range from −0.28 to +0.74 and because we are not focusing on the behavior of the model at extreme values. We log-transformed the frequency predictor because it is positively skewed, resulting in a clearer linear relationship with Lee's *L*.

The model is summarized in **Table 2**. All individual predictors in the fixed-effects component of our model are significant, while the variance component of our model indicates that a substantial

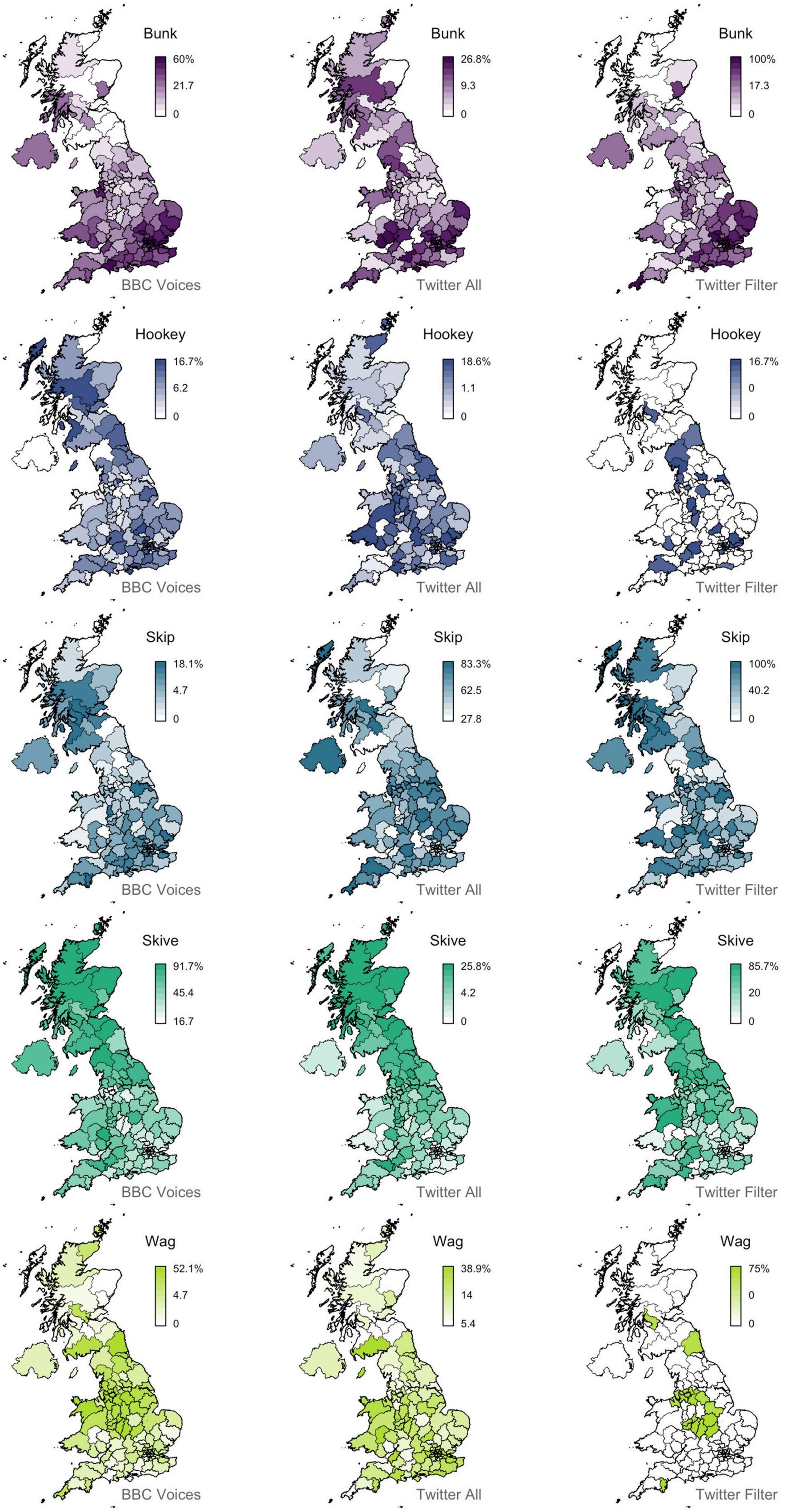


FIGURE 6 | *Bunk/hookey/skip/skive/wag* alternation comparison.

TABLE 3 | Descriptive statistics for the *playing truant* variants before and after filtering.

	Variant	*Corpus frequency*	*Spatial clustering*: Moran's *I*	*Polysemy*: percentage of target uses	*Map alignment*: Lee's *L*
All tokens	*Bunk*	4757	0.39	28	0.47
	Hookey	808	0.10	10	−0.04
	Skip	28272	0.19	2	−0.13
	Skive	2666	0.54	82	0.52
	Wag	7549	0.21	0	−0.06
Filtered tokens	*Bunk*	559	0.49	100	0.57
	Hookey	41	0.11	100	0.07
	Skip	985	0.13	100	0.00
	Skive	547	0.38	100	0.39
	Wag	49	0.20	100	0.33

amount of variability in Lee's *L* is attributable to variation across the 36 alternations. As expected, Moran's *I* and the percentage of target meanings are positively correlated with Lee's *L*, indicating that Twitter maps tend to be better matches when they show clear regional patterns and when they are primarily based on occurrences of word tokens with the target meaning. Frequency, however, is negatively associated with Lee's *L*, indicating that Twitter maps tend to be better matches when they are based on fewer tokens. This result is surprising. Although it suggests that our corpus is large enough to investigate this set of alternations, we believe that it also likely points to a fundamental issue with the ability of dialect surveys, as opposed to Twitter corpora, to map common words that are in use across the region of interest, often in alternation with less common regional words in the language of individuals. The relative usage of such words can still show continuous regional patterns, but it is difficult for such patterns to be mapped using surveys, where informants generally report one word per question. The drop in alignment as frequency rises may therefore reflect inaccuracies in the survey maps for common words, as opposed to the Twitter maps.

Finally, we can use our model to propose some guidelines about how likely new Twitter maps are to generalize—without taking survey data, which is rarely available, into consideration. These guidelines are useful because they allow dialectologists who map regional lexical variation using Twitter corpora to assess how confident they should be that their maps identify general patterns. For example, if one is interested in mapping general dialect regions through the multivariate analysis of Twitter lexical alternation maps, these guidelines could be used to filter out maps that are less likely to generalize, prior to aggregation. **Figure 5** illustrates how the expected value of Lee's *L* for map pairs changes as a function of the Moran's *I* and target token percentage, when log-transformed frequency takes its mean value. The solid and dashed lines represent cut-off values for Lee's *L* of 0.15 and 0.40 and were drawn to facilitate the assessment of the reliability of the alignment with a given combination of predictor values. For example, if we take a Lee's *L* value of 0.15 as being indicative of alignment, Twitter maps that have a Moran's *I* of at least 0.35 and are based on at least 50% target meanings can be expected to generalize.

Dealing With Polysemy

As is common in Twitter dialect studies, we did not control for polysemy (and homophony). We found, however, that high levels of polysemy do affect the generalizability and presumably by extension the accuracy of these maps. To deal with this issue, methods for word sense disambiguation can be applied. At the most basic level, all the tokens of the relevant forms can be hand-coded. This is most accurate, but it is an extremely time-consuming task and thus usually impractical when working with large corpora or feature sets. Alternatively, various more advanced approaches could be applied. For example, a sample of tokens can be hand-coded and then a machine learning classifier can be trained on this data and used to code other tokens (Austen, 2017), or a token-based semantic vector space model could be applied (Hilpert and Saavedra, 2017). A simpler and more transparent approach is to only count tokens that occur in contexts where the target meaning is especially likely.

For example, as summarized in the first half of **Table 3**, the *playing truant* alternation, which includes 5 variants, shows considerable polysemy in our Twitter corpus, based on our hand coding of 50 random tokens of the form drawn from our corpus. Only *skive,* which is the variant with the best alignment, occurs with its target meaning over 50% of the time. The only other variant with a strong alignment is *bunk,* which remarkably occurs with its target meaning only 28% of the time, illustrating how a regional signal can be detected even when the target meaning is relatively rare. The other three variants, however, occur with their target meanings at most 10% of the time and show negative alignments, making them three of the worst matches in the feature set. Notably, the strength of alignment is clearly associated with the amount of spatial clustering, but there is no clear relationship with frequency. For example, *hookey,* which is the most infrequent variant, shows poor alignment, but so does *skip*, which is by far the most frequent variant.

To test whether we can improve the maps for this alternation through simple word-sense disambiguation we recounted these variants in the Twitter corpus in restricted contexts, identified based on concordance line analysis. Specifically, we only counted tokens of *skip* when it was immediately followed by *class, classes,* *college, lecture, school, uni, university,* or *work; bunk*, *skive,* and *wag* when followed by these words or *off*; and *hookey* when preceded by a form of the verb *play*. We then recomputed the variant percentages, as well as the three map characteristics used as predictors of our model. The results are presented in the second half of **Table 3**, while the variants in all three datasets are mapped in **Figure 6**.

Overall, there is a substantial rise in alignment after filtering: all three variants with negative correlations now show positive correlations, most notably *wag*. We also see a clear improvement in the alignment for *bunk*. Alternatively, although the alignment is still strong, we see a decrease for *skive*, presumably because the number of tokens examined has been drastically reduced, even though the vast majority of these tokens were used with the target meaning. This highlights the main limitation with word sense disambiguation: in most cases it will greatly reduce token counts, potentially down to problematic levels. For example, consider the maps for the rarest of these words: after filtering there are very few tokens left for *hookey* and *wag*, resulting in maps where most areas have no attestation at all of the form, suggesting that the corpus is too small to map these variants. Nevertheless, as the map for *wag* illustrates, such maps can still represent improvements over the unfiltered versions in terms of alignment with the survey data.

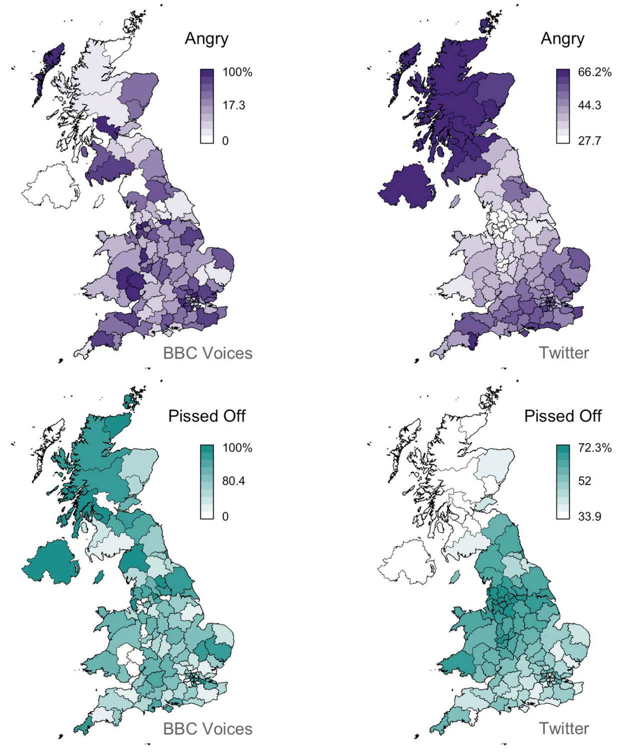


FIGURE 7 | *Angry/pissed off* alternation.

DISCUSSION

Although Twitter corpora are increasingly being used as the basis for dialect maps, their generalizability had not been established. Do these maps tell us anything about general patterns of regional variation, including in the spoken vernacular? Can these maps extend our general understanding of language variation and change? These are important questions because currently Twitter is the only data source from which precisely geolocated texts can be sampled at scale. Twitter maps have the potential to answer a range of basic questions in regional dialectology, but only if they are generalizable. In this study, we therefore set out to systematically test if Twitter maps, based on a 1.8 billion word corpus of geolocated Tweets collected in 2014 from across the UK, align with traditional survey maps, based on an unbiased sample of 139 lexical dialect maps taken from the BBC Voices dialect survey. Overall, we found broad correspondence between the two datasets, with a majority of the 139 map pairs showing meaningful levels of alignment in our opinion. In most cases, these two sets of maps agree across the four nations of the UK and within England between the North, the Midlands, and the South, although a substantial number of map pairs show more precise correspondence, for example identifying specific cities as hotspots for certain words. Given how different these two approaches to data collection are, we believe the alignment between these maps is strong evidence that Twitter maps are able to identify general dialect patterns.

The main outcome of this study is therefore validating the use of Twitter corpora for the analysis of general patterns of regional lexical variation, at least in British English. This is an important result for regional dialectology, because there are many advantages to working with dialect corpora as opposed to dialect surveys. Not only is it far easier to build corpora than conduct surveys, but dialect corpora allow for the open-ended analysis of a far wider range of features than surveys, which can only be used to collect data on a limited number of pre-selected features. Corpora also generally improve the resolution of dialect maps, allowing for more informants to be sampled over more locations. For example, our Twitter corpus contains posts from 1.9 million accounts, whereas the BBC Voices dataset contains responses from 84,000 informants. Finally, the fundamental reason to prefer dialect corpora is that they allow patterns of regional variation to be observed in natural language, whereas surveys only provide the opportunity to observe the linguistic opinion of informants, elicited in a single and very artificial communicative context.

For all these reasons, we believe that Twitter corpora can be the basis for general inquiry into regional lexical variation. However, we also believe that our analysis suggests that Twitter maps may generally provide a better foundation for dialectology than survey data, allowing for regional patterns to be identified more accurately in many cases. Perhaps the most striking example is the alternation between *angry* and *pissed off*, which is mapped in **Figure 7**. The Twitter maps identify much stronger regional patterns than the survey maps for these two variants, especially for *angry,* which shows limited spatial clustering in the survey data (Moran's $I = 0.10$), but a clear pattern in the Twitter data (Moran's $I = 0.80$). This example not only demonstrates how common words like *angry,* which are in usage across the UK, can show regional patterns and how these patterns can be identified through corpus analysis, but that such patterns can be difficult to access through surveys. This is reflected by the fact that the BBC Voices data for *angry* range from 0 to 100%, indicating that in some postal code areas no informant provided *angry*, whereas the Twitter analysis finds that in no postal code is either variant used <28% of the time. This result appears to expose a major limitation with standard survey-based approached to data collection in dialectology: individual informants can usually only supply a single variant per question, even when the informant uses multiple variants in their daily lives. In such cases, the maps for these variants, especially standard forms like *angry* that are clearly commonly used across the entire region of interest, may not accurately reflect patterns of regional linguistic variation in the population. The Twitter maps therefore seem to be more realistic than the survey maps, and by extension more reliable, although further research is necessary to directly test this hypothesis.

In addition to offering important validation for corpus-based approaches to regional dialectology, this study makes several other methodological contributions to the field. Perhaps of greatest value, we provide general quantitative guidelines for judging if Twitter-based maps are likely to generalize. We also introduce a new method for map comparison, Lee's L, which we borrowed from spatial analysis and which provides a more principled method for map correlation than approaches currently used in dialectology. We also show, however, that map comparison based on non-spatial correlation analysis yields similar results, offering support for the long tradition in dialectometry of using what are essentially correlation-based methods for aggregation (like Factor Analysis and Principal Components Analysis). Although we found Twitter maps to be remarkably robust in the face of polysemy, we also began to explore the use of techniques for word sense disambiguation to improve the reliability of lexical dialect maps; there is considerably more work to be done in this area. In addition, while we believe our results show that corpus-based approaches to dialectology are at least as powerful as survey-based approaches, our results also offer support for the generalisability of dialect surveys, whose validity has long been questioned, especially from outside the field (e.g., Pickford, 1956).

Descriptively, this study also presents one of the few corpus-based analyses of regional variation on the national level in modern British English. British dialectologists have not fully engaged with methods from computational sociolinguistics, and research has thus progressed slowly in recent years compared to American English. Consequently, there is much less agreement on issues such as the modern dialect regions of the UK than in the US, or how these regions are changing over time. These are the types of basic questions that British dialectologists can now pursue through the analysis of Twitter corpora, confident their results can provide insights about general patterns of regional linguistic variation in the UK.

Furthermore, our results not only offer evidence of the general value of Twitter corpora for theoretical research in dialectology, but they are themselves of direct relevance to our understanding of regional linguistic variation and change. Our main finding in this regard is that patterns of regional lexical variation are relatively stable across data sources—at least sufficiently stable for broad patterns of regional lexical variation to align. This result implies that patterns of regional lexical variation are relatively stable across communicative contexts. In fact, we find considerable evidence that the alternations behave quite differently in these two datasets: the median absolute difference in the maximum percentage of the 139 variants in the two datasets is 27%. In part, this is because of differences in how lexical alternation was measured, but the differences are so dramatic that it seems reasonable to assume that context matters in this regard. For example, the map for *bairn* (see **Figure 2**) shows that the variant is returned by up to 100% of informants in some areas the BBC Voices survey, but never accounts for more than 7% of the tokens of this alternation in any area in our Twitter corpus. Despite such differences in scale, these two maps show good alignment overall ($L = 0.43$). This result presumably obtains because the effect of situational variation is relatively consistent across the region: the percentage of *bairn* in the Twitter corpus drops dramatically, but the magnitude of this drop is relatively similar across the map, resulting in the same basic regional pattern being found in both datasets.

We believe this is an important result that sheds light on the relationship between the regional and situational determinants of language variation and change—an area that has been largely overlooked in dialectology and sociolinguistics, at least in part because dialect surveys and sociolinguistic interviews do not allow for situational variation to be analyzed in detail, as they involve eliciting data in one very specific and artificial context. Of course, there is still considerable disagreement between the two sets of maps, and our analysis of various characteristics of the Twitter maps only accounted for a proportion of this misalignment. Some of this variation may well be due to the interplay between region and situation. For example, it may be the case that people in different regions are using Twitter for a quantitatively different range of communicative purposes. Further research is necessary to explore these relationships, including analyzing and comparing regional variation in corpora representing other varieties of natural language, which will increasingly become possible as more and more language data comes online. However, much of this misalignment may also be due to social factors, which we have not considered in this study. In particular, we know that the demographics of our Twitter corpora do not match the demographics of the general population or presumably of the informants who responded to the BBC Voices survey. Similarly, some of this misalignment may be explained by our choice not to filter our Twitter dataset, for example by removing re-tweets. Our goal here was to evaluate the baseline level of alignment between Twitter dialect corpora and dialect surveys. How this alignment can be improved through more complex approaches to corpus construction could be the focus of future research now that we have set a baseline level of alignment.

Unfortunately, the analysis of social variation in Twitter is nowhere near as straightforward as the analysis of regional variation at this time, as the requisite metadata is not recorded or provided by Twitter or other social media platforms. Increasingly, however, researchers are developing powerful methods for estimating the demographics of Twitter users, based on a wide range of factors (e.g., Wang et al., 2019). Furthermore, there can be little doubt that as more and more of our lives are played out online increasing amounts of detailed social metadata will become available to researchers, as well as increasing amount of language data from across a wide range of registers, including the spoken vernacular. This will transform how we conduct sociolinguistic research. To truly understand how language variation and change functions as a system, across region, society, and communicative contexts, we must adopt a corpus-based approach to data collection. This is the only way that variation can be observed in a wide range of linguistic variables across a wide range of social and situational contexts. This is the promise of computational sociolinguistics and the future of our field.

DATA AVAILABILITY

All datasets and code used for this study are included in the manuscript and/or the **Supplementary Files**.

AUTHOR CONTRIBUTIONS

JG, CM, and AN contributed conception and design of the study and wrote the first draft of the manuscript. DG, CM, JG, and AN organized the database. JG, AM, and AN conducted statistical analysis. JG, CM, AN, and AM wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was funded by the Arts and Humanities Research Council (UK), The Economic and Social Research Council (UK), Jisc (UK), and the Institute of Museum and Library Services (US), as part of the Digging into Data Challenge (Round 3).

ACKNOWLEDGMENTS

We would like to thank Tam Blaxter, Gabriel Doyle, Jonathan Dunn, Zhaoya Gong, Deepthi Gopal, Jason Grafmiller, Alice Kasakoff, John Nerbonne, Benedikt Szmrecsanyi, Emily Waibel, David Willis, and Bodo Winter for their comments on this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2019.00011/full#supplementary-material

REFERENCES

Anderwald, L. (2009). *The Morphology of English Dialects: Verb-Formation in Non-standard English*. Cambridge, UK: Cambridge University Press.

Asprey, E. (2007). *Black Country English and Black Country Identity* (Unpublished PhD thesis). University of Leeds.

Austen, M. (2017). "Put the groceries up": comparing black and white regional variation. *Am. Speech* 92, 298–320. doi: 10.1215/00031283-4312064

Bailey, G. (2015). "Orthographic reflections of (ing): a Twitter-based corpus study," *Paper Presented at Manchester Forum in Linguistics* (Manchester: University of Manchester).

Bailey, G. (2016). "Regional variation in 140 characters: mapping geospatial tweets," *Paper Presented at Workshop on Using Twitter for Linguistic Research* (Canterbury: University of Kent).

Bishop, H., Coupland, N., and Garrett, P. (2005). Conceptual accent evaluation: thirty years of accent prejudice in the UK. *Acta Linguist. Hafniensia* 37, 131–154. doi: 10.1080/03740463.2005.10416087

Brook, G. L. (1963). *English Dialects*. London: Deutsch.

Burbano-Elizondo, L. (2008). *Language variation and identity in Sunderland* (Unpublished PhD thesis). University of Sheffield.

Cook, P., Han, B., and Baldwin, T. (2014). Statistical methods for identifying local dialectal terms from GPS-tagged documents. *Dictionaries* 35, 248–271. doi: 10.1353/dic.2014.0020

Doyle, G. (2014). "Mapping dialectal variation by querying social media," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, eds S. Wintner, S. Goldwater, and S. Riezler (Gothenburg), 98–106.

Durham, M. (2016). "Changing attitudes towards the welsh english accent: a view from Twitter," in *Sociolinguistics in Wales*, eds M. Durham and J. Morris (Basingstoke: Palgrave), 181–205.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2012). Mapping the geographical diffusion of new words. *PLOS ONE* 9.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE* 9:e113114. doi: 10.1371/journal.pone.0113114

Elmes, S. (2013). "Voices: a unique BBC adventure," in *Analysing 21st Century British English: Conceptual and Methodological Aspects of the "Voices" Project*, eds C. Upton and B. Davies (London: Routledge), 1–11.

Grieve, J. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English* (Ph.D. dissertation). Northern Arizona University.

Grieve, J. (2013). A statistical comparison of regional phonetic and lexical variation in American English. *Lit. Linguist. Comput.* 28, 82–107. doi: 10.1093/llc/fqs051

Grieve, J. (2016). *Regional Variation in Written American English*. Cambridge, UK: Cambridge University Press.

Grieve, J. (2017). "Assessing smoothing parameters in dialectometry," in *From Semantics to Dialectometry: Festschrift in Honor of John Nerbonne*, eds M. Wieling, M. Kroon, G. van Noord, and G. Bouma (Tributes 32, College Publications), 119–126.

Grieve, J. (2018). "Spatial statistics for dialectology," in *The Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Oxford: Wiley-Blackwell), 415–433.

Grieve, J., Asnaghi, C., and Ruette, T. (2013). Site-restricted web searches for data collection in regional dialectology. *Am. Speech* 88, 413–440. doi: 10.1215/00031283-2691424

Grieve, J., Nini, A., and Guo, D. (2017). Analyzing lexical emergence in Modern American English online. *Engl. Lang. Linguist.* 21, 99–127. doi: 10.1017/S1360674316000113

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on American social media. *J. Engl. Linguist.* 46, 293–319. doi: 10.1177/0075424218793191

Hilpert, M., and Saavedra, D. C. (2017). Using token-based semantic vector spaces for corpus-linguistic analyses: from practical applications to tests of theoretical. *Corpus Linguist. Linguist. Theory*. 1–32. doi: 10.1515/cllt-2017-0009

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Comput. Environ. Urban Syst.* 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Ihalainen, O., Kyto, M., and Rissanen, M. (1987). "The Helsinki corpus of english texts: diachronic and dialectal report on work in progress," in *Corpus Linguistics and Beyond, Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam, 21–32.

Jones, G. E., Robert, O. J., Alan, R. T., and David, T. (2000). *The Welsh Dialect Survey*. Cardiff: University of Wales Press.

Jones, T. (2015). Toward a description of African American vernacular english dialect regions using "Black Twitter". *Am. Speech* 90, 403–440. doi: 10.1215/00031283-3442117

Kulkarni, V., Perozzi, B., and Skiena, S. (2016). "Freshman or fresher? Quantifying the geographic variation of internet language," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, eds M. Strohmaier and K. P. Gummadi (Palo Alto, CA: The AAAI Press), 615–618.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Philadelphia Press.

Lee, S.-I. L. (2001). Developing a bivariate spatial association measure: an integration of Pearson's *r* and Moran's *I. J. Geogr. Syst.* 3, 369–385. doi: 10.1007/s101090100064

Leemann, A., Marie-José, K., and David, B. (2018). The English Dialects App: the creation of a crowdsourced dialect corpus. *Ampersand* 5, 1–17. doi: 10.1016/j.amper.2017.11.001

Llamas, C. (1999). A new methodology: data elicitation for social and regional language variation studies. *Leeds Work. Pap. Linguist. Phon.* 7, 95–118.

Llamas, C. (2007). "A place between places": language and identities in a border town. *Lang. Soc.* 36, 579–604. doi: 10.1017/S0047404507070455

Longley, P. A., Adnan, M., and Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environ. Plann. A* 47, 465–484. doi: 10.1068/a130122p

MacKenzie, L., Bailey, G., and Danielle, T. (2015). *Our Dialects: Mapping Variation in English in the UK*. Available online at: http://tiny.cc/OurDialects

Maguire, W. (2012). Mapping the existing phonology of english dialects. *Dialectol. Geolinguist.* 20, 84–107. doi: 10.1515/dialect-2012-0006

Mather, J. Y., Speitel, H. H., and Leslie, G. W. (1975). (eds.). *The Linguistic Atlas of Scotland, Scots Section, Vol. 1*. Hamden, CT: Archon Books.

Mellon, J., and Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Res Polit*. 4, 1–9. doi: 10.1177/2053168017720008

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Lit. Linguist. Comput.* 21, 463–475. doi: 10.1093/llc/fql041

Nguyen, D., Dogruöz, A. S., Ros,é, C. P., and De Jong, F. (2016). Computational sociolinguistics: a survey. *Comput. Linguist.* 42, 537–593. doi: 10.1162/COLI_a_00258

Nini, A., Corradini, C., Guo, D., and Grieve, J. (2017). The application of growth curve modeling for the analysis of diachronic corpora. *Lang. Dyn. Change.* 7, 102–125. doi: 10.1163/22105832-00701001

O'Dochartaigh, C. (1994). *Survey of the Gaelic Dialects of Scotland: Questionnaire Materials Collected for the Linguistic Survey of Scotland*. Dublin: Dublin Institute for Advanced Studies; School of Celtic Studies.

Orton, H. (1962). *Survey of English dialects: Introduction*. Leeds: Arnold.

Parry, D. (1999). *A Grammar and Glossary of the Conservative Anglo-Welsh Dialects of Rural Wales*. Sheffield: NATCECT.

Pickford, G. R. (1956). American linguistic geography: a sociological appraisal. *Word* 12, 211–233. doi: 10.1080/00437956.1956.11659600

Rahimi, A., Cohn, T., and Baldwin, T. (2017). A Neural model for user geolocation and lexical dialectology. *arXiv*. 209–216. doi: 10.18653/v1/P17-2033

Robinson, J., Herring, J., and Gilbert, H. (2013). "The British library description of the BBC voices recordings collection," in *Analysing 21st Century British English: Conceptual and Methodological Aspects of the "Voices" Project, 1st Edn*, eds C. Upton and B. Davies (London; New York, NY: Routledge), 136–161.

Shackleton, R. (2007). Phonetic variation in the traditional English dialects: a computational analysis. *J. Engl. Linguist.* 35, 30–102. doi: 10.1177/0075424206297857

Sheidlower, J. (2018). *The Closing of a Great American Dialect Project*. The New Yorker. Available online at: https://www.newyorker.com/culture/cultural-comment/the-closing-of-a-great-american-dialect-project (accessed September 22, 2017).

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., and Goldwater, S. (2017). "Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1239–1248. (Valencia: Association for Computational Linguistics). doi: 10.18653/v1/E17-1116

Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.

Trevisa, J. (1495) *Policronicon*. Westminster: Wynkyn Theworde.

Upton, C. (2013). "Blurred boundaries: the dialect word from the BBC," in *Analysing 21st Century British English: Conceptual and Methodological Aspects of the "Voices" Project*, eds C. Upton and B. Davies (London: Routledge), 180–197.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flack, F., et al. (2019). "Demographic inference and representative population estimates from multilingual social media data," in *Proceeding of WWW '19 The World Wide Web Conference* (San Francisco, CA: ACM), 2056–2067.

Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geogr. Anal.* 17, 263–283. doi: 10.1111/j.1538-4632.1985.tb00849.x

Wieling, M., Upton, C., and Thompson, A. (2014). Analyzing the BBC voices data: contemporary english dialect areas and their characteristic lexical variants. *Lit. Linguist. Comput.* 29, 107–117. doi: 10.1093/llc/fqt009

Willis, D., Leemann, A., Gopal, D., and Blaxter, T. (2018). "Localising morphosyntactic variation in Welsh Twitter data," *Presented at NWAV 47* (New York, NY).

Wright, J. (ed.). (1898). *The English Dialect Dictionary: A-C, Vol. 1*. Oxford: Oxford University Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 14 August 2019
doi: 10.3389/frai.2019.00015

Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology

*Jonathan Dunn**

Department of Linguistics, University of Canterbury, Christchurch, New Zealand

The goal of this paper is to provide a complete representation of regional linguistic variation on a global scale. To this end, the paper focuses on removing three constraints that have previously limited work within dialectology/dialectometry. First, rather than assuming a fixed and incomplete set of variants, we use Computational Construction Grammar to provide a replicable and falsifiable set of syntactic features. Second, rather than assuming a specific area of interest, we use global language mapping based on web-crawled and social media datasets to determine the selection of national varieties. Third, rather than looking at a single language in isolation, we model seven major languages together using the same methods: Arabic, English, French, German, Portuguese, Russian, and Spanish. Results show that models for each language are able to robustly predict the region-of-origin of held-out samples better using Construction Grammars than using simpler syntactic features. These global-scale experiments are used to argue that new methods in *computational sociolinguistics* are able to provide more generalized models of regional variation that are essential for understanding language variation and change at scale.

Keywords: dialectology, dialectometry, construction grammar, syntactic variation, text classification, language mapping, dialect mapping, computational sociolinguistics

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Benjamin Luke Johnson,
University of Maryland, Baltimore,
United States
Melanie Röthlisberger,
University of Zurich, Switzerland

****Correspondence:***
Jonathan Dunn
jonathan.dunn@canterbury.ac.nz

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

***Received:** 21 April 2019*
***Accepted:** 22 July 2019*
***Published:** 14 August 2019*

Citation:
Dunn J (2019) Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology.
Front. Artif. Intell. 2:15.
doi: 10.3389/frai.2019.00015

1. INTRODUCTION

This paper shows that computational models of syntactic variation provide precise and robust representations of national varieties that overcome the limitations of traditional survey-based methods. A computational approach to variation allows us to systematically approach three important problems: First, what set of variants do we consider? Second, what set of national dialects or varieties do we consider? Third, what set of languages do we consider? These three questions are usually answered in reference to the *convenience* or *interests* of the research project at hand. From that perspective, the goal of this paper is global, multi-lingual, whole-grammar syntactic dialectometry. Previous work has performed whole-grammar dialectometry with Construction Grammars, first using a pre-defined inventory of national varieties (Dunn, 2018a) and then using data-driven language mapping to select the inventory of national varieties (Dunn, 2019b). This paper further extends computational dialectometry by studying seven languages across both web-crawled and social media corpora. The paper shows that a classification-based approach to syntactic variation produces models that (i) are able to make accurate predictions about the region-of-origin of held-out samples, (ii) are able to characterize the aggregate syntactic similarity between varieties, and (iii) are able to measure the uniqueness of varieties as an empirical correlate for qualitative notions like *inner-circle* vs. *outer-circle*.

What features do we use for dialectometry? Most previous work relies on phonetic or phonological features (Kretzschmar, 1992, 1996; Heeringa, 2004; Labov et al., 2005; Nerbonne, 2006, 2009; Grieve et al., 2011, 2013; Wieling and Nerbonne, 2011, 2015; Grieve, 2013; Nerbonne and Kretzschmar, 2013; Kretzschmar et al., 2014; Kruger and van Rooy, 2018) for the simple reason that phonetic representations are relatively straight-forward: a vowel is a vowel and the measurements are the same across varieties and languages. Previous work on syntactic variation has focused on either (i) an incomplete set of language-specific variants, ranging from only a few features to hundreds (Sanders, 2007, 2010; Szmrecsanyi, 2009, 2013, 2014; Grieve, 2011, 2012, 2016; Collins, 2012; Schilk and Schaub, 2016; Szmrecsanyi et al., 2016; Calle-Martin and Romero-Barranco, 2017; Grafmiller and Szmrecsanyi, 2018; Tamaredo, 2018) or (ii) language-independent representations such as function words (Argamon and Koppel, 2013) or sequences of part-of-speech labels (Hirst and Feiguina, 2007; Kroon et al., 2018). This forces a choice between either an *ad hoc* and incomplete syntactic representation or a reproducible but indirect syntactic representation.

This previous work on syntactic dialectometry has depended on the idea that a grammar is an inventory of specific structures: the double-object construction vs. the prepositional dative, for example. Under this view, there is no language-independent feature set for syntax in the way that there is for phonetics. But we can also view syntax from the perspective of a discovery-device grammar (Chomsky, 1957; Goldsmith, 2015): in this case, our theory of grammar is not a specific description of a language like English but rather a function for mapping between observations of English and a lower-level grammatical description of English: $G = D(\text{CORPUS})$. Thus, a discovery-device grammar (G) is an abstraction that represents what the grammatical description would be if we applied the learner (D) to a specific sample of the language (CORPUS). A discovery-device grammar allows us to generalize syntactic dialectometry: we are looking for a model of syntactic variation, V, such that when applied to a grammar, $V(G)$, the model is able to predict regional variation in the grammar. But G is different for each language, so we generalize this to $V(D(\text{CORPUS}))$. In other words, we use an independent corpus for each language as input to a discovery-device grammar and then use the resulting grammar as a feature space for performing dialectometry. This approach, then, produces an inventory of syntactic features for each language in a reproducible manner in order to replace hand-crafted syntactic features. The specifics of the datasets used for modeling regional variation are described in section 2.1 and the discovery-device grammar used to create reproducible feature sets is described in section 2.2.

What type of model should we use to represent global syntactic variation? Previous work has relied largely on unsupervised methods like clustering (Wieling and Nerbonne, 2011), factor analysis of spatial autocorrelation scores (Grieve, 2013), and individual differences scaling as an extension of multidimensional scaling (Ruette and Speelman, 2014). These models attempt to aggregate individual variants into larger bundles of features: which individual features represent robust aggregate isoglosses with a similar geographic extent? The problem is that it is difficult to evaluate the predictions of one such bundle against another. While useful for visualizations, these models are difficult to evaluate against ground-truths. Another strand of work models the importance of predictor variables on the use of a particular variant, with geographic region as one possible predictor (Szmrecsanyi et al., 2016). These models are based on multivariate work in sociolinguistics that attempts to find which linguistic, social, or geographic features are most predictive of a particular variant.

While useful for understanding individual variants, however, these models are unable to handle the aggregation of variants directly. For example, although it is possible to create a distance matrix between regions for each individual feature and then to aggregate these matrices, the resulting aggregations are subject to variability: What is the best aggregation method? If two methods provide different maps, which should we prefer? How stable are aggregations across folds? On the one hand, we want dialectometry to establish a ground-truth about the regional distribution of variants and dialects. But, on the other hand, because unsupervised methods like clustering are subject to such potential variability, we also need a ground-truth to evaluate which aggregation method is the most accurate.

One solution to this problem is to take a classification approach, in which the ground-truth is the region-of-origin for individual samples. Given a model of dialectal variation, how accurately can that model predict the region-of-origin of new samples? For example, the idea is that a more complete description of the syntactic differences between Australian English and New Zealand English will be able to predict more accurately whether a new sample comes from Australia or New Zealand. This prediction task provides a ground-truth for aggregation. But it comes with two important caveats: First, a high prediction accuracy does not guarantee that the model captures all relevant variation, only that it captures enough variation to distinguish between national varieties. This can be mitigated, however, by using cross-fold validation and unmasking as shown in section 3.2. Second, while most work in dialectometry tries to establish geographic boundaries, this work assumes geographic boundaries (i.e., polygons of nation-states).

What languages and regions need to be represented in dialectometry? Because of coloniziation and globalization (Kachru, 1990), a few languages like English are now used around the world by diverse national communities. Even though these international languages have global speech communities, dialectology and sociolinguistics continue to focus largely on sub-national dialects, often within so-called *inner-circle* varieties (Kachru, 1982). This paper joins recent work in taking a global approach by using geo-referenced texts (Goldhahn et al., 2012; Davies and Fuchs, 2015; Donoso and Sanchez, 2017) to represent national varieties (Szmrecsanyi et al., 2016; Calle-Martin and Romero-Barranco, 2017; Cook and Brinton, 2017; Rangel et al., 2017; Dunn, 2018a, 2019b; Tamaredo, 2018). The basic point is that in order to represent regional variation as a complete system, dialectometry must take a global perspective. This paper uses data-driven language mapping to choose (i) which international languages are used widely enough to justify inclusion and (ii) which languages in which countries need to be included as

TABLE 1 | Size of geo-referenced corpora in words by region.

Region	Countries	Population	(%)	Web	(%)	Twitter	(%)
Africa, North	9	250 mil	3.4%	123.85 mil	0.7%	85.55 mil	2.1%
Africa, Southern	4	75 mil	1.0%	59.07 mil	0.4%	87.34 mil	2.1%
Africa, Sub-Saharan	73	742 mil	10.1%	424.75 mil	2.6%	254.20 mil	6.1%
America, Brazil	1	206 mil	2.8%	218.11 mil	1.3%	118.13 mil	2.9%
America, Central	25	214 mil	2.9%	886.61 mil	5.3%	383.81 mil	9.3%
America, North	2	355 mil	4.8%	236.59 mil	1.4%	350.12 mil	8.5%
America, South	11	210 mil	2.9%	1,163.00 mil	7.0%	402.15 mil	9.7%
Asia, Central	10	198 mil	2.7%	965.09 mil	5.8%	102.79 mil	2.5%
Asia, East	8	1,635 mil	22.3%	2,201.86 mil	13.2%	95.70 mil	2.3%
Asia, South	7	1,709 mil	23.3%	448.23 mil	2.7%	331.19 mil	8.0%
Asia, Southeast	22	615 mil	8.4%	2,011.06 mil	12.1%	245.18 mil	5.9%
Europe, East	17	176 mil	2.4%	4,553.10 mil	27.4%	322.46 mil	7.8%
Europe, Russia	1	144 mil	2.0%	101.44 mil	0.6%	105.04 mil	2.5%
Europe, West	25	421 mil	5.7%	2,422.85 mil	14.6%	823.80 mil	19.9%
Middle East	15	334 mil	4.5%	660.73 mil	4.0%	222.98 mil	5.4%
Oceania	8	59 mil	1.0%	164.02 mil	1.0%	213.06 mil	5.1%
Total	199	7.35 bil	100%	16.65 bil	100%	4.14 bil	100%

national varieties. We use geo-referenced corpora drawn from web pages and social media for both tasks. Seven languages are selected for dialectometry experiments: Arabic, English, French, German, Portuguese, Russian, and Spanish. These seven languages account for 59.25% of the web-crawled corpus and 74.67% of the social media corpus. The corpora are regionalized to countries. Thus, the assumption is that any country which frequently produces data in a language has a national variety of that language. For example, whether or not there is a distinct variety of New Zealand English depends entirely on how much English data is observed from New Zealand in these datasets. The models then have the task of determining how distinct New Zealand English is from other national varieties of English.

First, we consider the selection of (i) languages and (ii) national varieties of languages (section 2.1) as well as the selection of a syntactic feature space (section 2.2). We then present the specifics of the experimental framework (section 2.3). Second, we compare prediction accuracies by language and feature set (section 3.1), in order to measure the quality of the models. Next, we evaluate the robustness of the models across rounds of feature pruning and the similarity of the models across registers in order to examine potential confounds (section 3.2). Having validated the models themselves, the next section examines regional accuracies and the similarities between national varieties (section 3.3). Finally, we develop measures for the syntactic uniqueness of each regional variety (section 3.4) and search for empirical correlates of concepts like *inner-circle* and *outer-circle* within this corpus-based approach (section 3.5). Third, we discuss two important issues: the application of different categorizations like *inner-circle vs. outer-circle* or *native vs. non-native* to these datasets (section 4.1) and the implications of a computational approach to dialectometry for sociolinguistics more broadly (section 4.2).

2. MATERIALS AND METHODS

2.1. Language Mapping and Dialectometry

We begin with data-driven language mapping: First, what languages have enough national varieties to justify modeling? Second, which national varieties should be included for each language? Third, which datasets can be used to represent specific national varieties and how well do these datasets represent the underlying populations? This paper depends on geo-referenced corpora: text datasets with meta-data that ties each document to a specific place. The size of both datasets by region is shown in **Table 1**, together with ground-truth population data from the UN (United Nations, 2017). The size of each region relative to the entire dataset is also shown: for example, 14.6% of the web corpus comes from Western Europe which accounts for only 5.7% of the global population. This comparison reveals the over-representation and under-representation of each region.

Data comes from two sources of digital texts: web pages from the Common Crawl[1] and social media from Twitter[2]. The Common Crawl data represents a large snapshot of the internet; although we cannot direct the crawling procedures, we are able to process the archived web pages from the perspective of a geo-referenced corpus. The author of each individual web page

[1] http://www.commoncrawl.org

[2] http://www.twitter.com

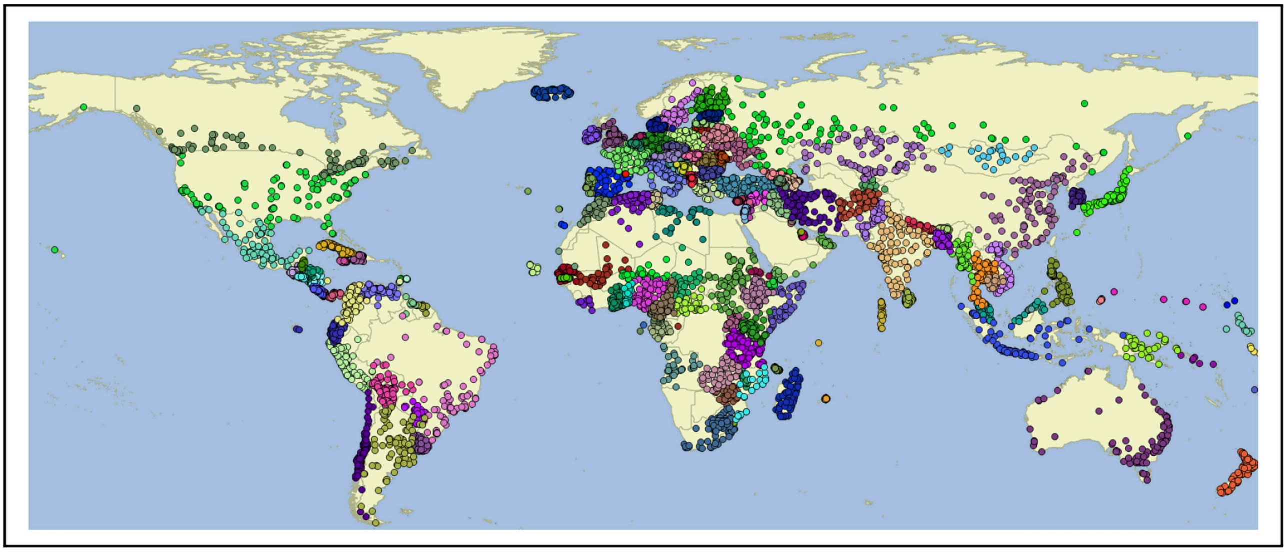


FIGURE 1 | Cities for Collection of Twitter Data (50 km radius from each).

may be unknowable but we can use country-specific top-level domains for country-level geo-referencing: for example, web pages under the *.nz* domain are from New Zealand. Previous work has shown that there is a relationship between domain-level geo-referenced web pages and national varieties (Cook and Brinton, 2017). Some countries are not available because their top-level domains are used for other purposes (i.e., *.ai*, *.fm*, *.io*, *.ly*, *.ag*, *.tv*). Domains that do not contain geographic information are also removed from consideration (e.g., *.com* sites). The Common Crawl dataset covers 2014 through the end of 2017, totalling 81.5 billion web pages. As shown in **Table 1**, after processing this produces a corpus of 16.65 billion words. This dataset represents 166 out of 199 total countries considered in this paper. Some countries do not use their country-level domains as extensively as others: in other words, *.us* does not account for the same proportion of web pages from the United States as *.nz* does from New Zealand. It is possible that this skews the representation of particular areas. Thus, **Table 1** shows the UN-estimated population for each region as reference. The web corpus is available for download[3] as is the code used to create the corpus[4].

In isolation, web-crawled data provides one observation of global language use. Another common source of data used for this purpose is Twitter [e.g., (Eisenstein et al., 2010, 2014; Roller et al., 2012; Kondor et al., 2013; Mocanu et al., 2013; Graham et al., 2014; Donoso and Sanchez, 2017)]. The shared task at PAN-17, for example, used Twitter data to represent national varieties of several languages (Rangel et al., 2017). A spatial search is used to collect Tweets from within a 50 km radius of 10 k cities[5]. This city-based search avoids biasing the selection by using language-specific keywords or hashtags. A map of each city used for collection is shown in **Figure 1**; while this approach avoids a language-bias, it could under-represent rural areas given the 50 km radius of each collection area. The Twitter data covers the period from May of 2017 until early 2019, drawn from the Twitter API using a spatial query. This creates a corpus containing 1,066,038,000 Tweets. The language identification component, however, only provides reliable predictions for samples containing at least 50 characters (c.f., the language id code[6] and the models used[7]). Thus, the corpus is pruned to include only Tweets above that length threshold. As shown in **Table 1**, this produces a corpus containing 4.14 billion words. While the Common Crawl corpus represents 166 countries, the Twitter corpus represents 169. There are 33 countries that only Twitter represents (not the Common Crawl) and 30 that only the Common Crawl represents (not Twitter). This shows the importance of drawing on two different sources of language use.

Given the idiosyncracies of these two datasets (i.e., the availability of country-codes for web data and the selection of cities for Twitter data), it is quite likely that each represents different populations or, at least, that each represents different registers of language usage from the same population. We can use ground-truth population data to deal with the problem of different populations. First, notice that both datasets under-represent all regions in Africa; but the web dataset has the worst under-representation: while Africa accounts for 14.5% of the world's population, it accounts for only 3.7% of the web corpus. The Americas and Europe, on the other hand, are over-represented in both datasets. Twitter especially over-represents North America (8.5% of the corpus vs. 4.8% of the population); but the web corpus under-represents North America (only 1.4%

[3] https://labbcat.canterbury.ac.nz/download/?jonathandunn/CGLU_v3

[4] https://github.com/jonathandunn/common_crawl_corpus

[5] https://github.com/datasets/world-cities

[6] https://github.com/jonathandunn/idNet

[7] https://labbcat.canterbury.ac.nz/download/?jonathandunn/idNet_models

of the corpus), mostly from the lack of adoption of the *.us* domain. Western Europe is over-represented in both corpora: while it acounts for only 5.7% of the population, it provides 14.6% of the web corpus and 19.9% of the Twitter corpus. Although these trends are expected, it is helpful to quantify the degree of over-representation. Less expectedly, the web corpus greatly over-represents Eastern Europe (27.4% of the corpus but only 2.4% of the population). Asia, especially the East and South, are under-represented in both datasets.

On the one hand, the use of population data here allows us to quantify exactly how each of these datasets is skewed. On the other hand, our purpose is to model regional syntactic variation: do the datasets need to be prefectly aligned with regional populations in order to achieve this? There are two observations to be made: First, if a region is over-represented then we do not need to worry about missing any national varieties from that area; but we should be worried about over-representing those particular national varieties (this is why there is a cap on the number of training samples from each dialect). Second, it could be the case that we are missing national varieties from under-represented areas. For example, any missing national varieties are likely to be from Africa or East Asia, given the skewed representations of this dataset. Related work, however, has shown that it in the case of major international languages like those considered here, the problem is over-representation rather than under-representation in the form of missing regional varieties (Dunn and Adams, 2019). We leave it to future work to make improvements in the selection of regional varieties using population-based sampling to overcome skewness in corpus distributions.

What languages should be included in a model of global syntactic variation? Given that we are using countries to define regional varieties, a language needs to occur in many countries. Here we use a threshold of 1 million words to say that a language is used significantly in a given country. **Table 2** shows the seven languages included in this study, encompassing 59.25% of the web corpus and 74.67% of the Twitter corpus. Some other languages occur in several countries in one dataset but not the other and so are not included. For example, Italian occurs in 17 countries in the web corpus but only 2 in the Twitter corpus; Indonesian occurs in 10 countries in the web corpus but only 3 countries in the Twitter corpus. Given that we model varieties using a classifier, we focus on those languages that have a sufficient number of national varieties to make classification a meaningful approach.

TABLE 2 | Above: number of countries and words by language and domain and Below: number of varieties and test samples by language and domain.

Language	Countries (Web)	Words (Web)	Countries (Twitter)	Words (Twitter)
Arabic (ara)	19	348,671,000	25	179,473,000
English (eng)	130	4,990,519,000	137	1,552,268,000
French (fra)	36	479,857,000	24	176,009,000
German (deu)	24	500,029,000	7	71,234,000
Portuguese (por)	14	431,884,000	22	199,080,000
Russian (rus)	37	1,361,331,000	9	126,834,000
Spanish (spa)	43	1,757,200,000	44	789,239,000
	% of Total:	59.25%	% of Total:	74.67%

Language	Varieties (Web)	N. Test (Web)	Varieties (Twitter)	N. Test (Twitter)
Arabic (ara)	4	14,685	7	15,537
English (eng)	14	66,476	14	64,208
French (fra)	13	46,562	4	12,130
German (deu)	7	35,240	2	7,722
Portuguese (por)	4	15,129	2	8,650
Russian (rus)	19	84,925	3	9,164
Spanish (spa)	17	84,093	17	76,653

2.2. Finding Syntactic Variants

This paper represents syntactic variants using a discovery-device Construction Grammar (CxG) that produces a CxG for each language given an independent corpus representing that language. CxG itself is a usage-based paradigm that views grammar as a set of overlapping constructions made up of slot-fillers defined by syntactic, semantic, and sometimes lexical constraints (Goldberg, 2006; Langacker, 2008). This paper draws on recent approaches to computational modeling of CxGs (Dunn, 2017, 2018b, 2019a), including previous applications of a discovery-device CxG to dialectometry for English (Dunn, 2018a, 2019b).

Constructions are represented as a sequence of slot-constraints, as in (1a). Slots are separated by dashes and constraints are defined by both type (Syntactic, Joint Semantic-Syntactic, Lexical) and by filler (for example: NOUN, a part-of-speech or ANIMATE, a semantic domain).

(1a) [SYN:NOUN — SEM-SYN:TRANSFER[V] — SEM-SYN:ANIMATE[N] — SYN:NOUN]
(1b) “He gave Bill coffee.”
(1c) “He gave Bill trouble.”
(1d) “Bill sent him letters.”
(2a) [SYN:NOUN — LEX:“give” — SEM-SYN:ANIMATE[N] — LEX:“a hand"]
(2b) “Bill gave me a hand.”

The construction in (1a) contains four slots: two with joint semantic-syntactic constraints and two with simple syntactic constraints. The examples in (1b) to (1d) are tokens of the construction in (1a). Lexical constraints, as in (2a), represent idiomatic sentences like (2b). A CxG is a collection of many individual constructions. For the purposes of dialectometry, these are quantified as one-hot encodings of construction frequencies. This, in essence, provides a bag-of-constructions that is evaluated against traditional bag-of-words features.

A large portion of the language-learning corpus for each language comes from web-crawled data (Baroni et al., 2009; Majliš and Žabokrtský, 2012; Benko, 2014) and data from the CoNLL 2017 Shared Task (Ginter et al., 2017). Because the goal is to provide a wide representation of each language, this is augmented by legislative texts from the EU and UN (Tiedemann, 2012; Skadiš et al., 2014), the OpenSubtitles corpus (Tiedemann, 2012), and newspaper texts. The exact collection of documents used for learning CxGs is available for download[8]. While both web-crawled and social media datasets are used to represent national varieties, the grammars used are learned mainly from web-crawled corpora. On the one hand, we use separate datasets for grammar learning and dialectometry in order to remove the possible confound that the grammars are over-fitting a specific dataset. On the other hand, we do not explicitly know which regional varieties the data used for grammar learning is drawn from. The discussion in section 3.5, as well as other work (Dunn, 2019b), shows that at least the English grammar better represents inner-circle varieties like UK English. In this case, then, we prefer to avoid the possible confound of over-fitting even though the result is a grammar that is learned from datasets implicitly drawn from inner-circle varieties.

This paper evaluates two alternate CxGs for dialectometry, alongside function words and lexical features: CxG-1 (Dunn, 2018a,b) and CxG-2 (Dunn, 2019a). As described and evaluated elsewhere (Dunn, 2019a), CxG-1 relies on frequency to select candidate slot-constraints while CxG-2 relies on an association-based search algorithm. The differences between the two competing discovery-device grammars as implementations of different theories of language learning are not relevant here. Rather, we evaluate both grammars because previous work (Dunn, 2018a) relied on CxG-1 and this comparison makes it possible to connect the multi-lingual experiments in this paper with English-only experiments in previous work. It should be noted, however, that other work has shown that association-based constraints out-perform frequency-based constraints across several languages (Dunn, 2019a). As shown in section 3, this paper finds that association-based constraints also perform better on the task of dialectometry. This is important because the evaluation connects the emergence of syntactic structure with variation in syntactic structure.

Previous work on syntactic dialectometry focuses on paired sets of features which can be viewed as alternate choices that express the same function or meaning. In other words, these approaches contrast constructions like the double object vs. the prepositional dative and then quantify the relative preference of particular varieties for one variant over the other. From our perspective, such an approach is essential for a limited feature space because syntactic variation is structured around different constructions that encode the same function or meaning. In other words, two constructions which have entirely different uses cannot be in competition with one another: contrasting the double object and the get-passive constructions, in isolation, is not a meaningful approach to syntactic variation because their frequencies are influenced by other unseen parts of the grammar. On the other hand, looking at the frequency of a single construction in isolation can be meaningful but will never reveal the full picture of syntactic variation.

This whole-grammar construction-based approach to dialectology represents as much of the functional space as possible. This provides an implicit pairing of syntactic variants: without a topic bias, we expect that the relative frequency of a specific construction will be consistent across documents. If one construction is more frequent, that indicates an increased preference for that construction. This approach does not explicitly pair variants because part of the problem is to learn which constructions are in alternation. From a different perspective, we could view alternating variants as knowledge that is traditionally given to models within quantitative sociolinguistics: which constructions are in competition with one another? But the idea here is to leave it to the model itself to determine which constructions are in competition.

Because this work is situated within both dialectometry and construction grammar, we view syntactic variation as fundamentally structured around function and meaning (as described above). But more traditional sociolinguistic and generativist work on syntactic variation does not share this underlying view. In this case the prediction task itself allows us to translate between competing assumptions: regardless of how we understand the source of variation, the models are ultimately evaluated on how well they are able to predict region-of-origin (samples from New Zealand vs. samples from Australia) using only syntactic representations. This type of ground-truth evaluation can be undertaken, with greater or lesser success, with any set of assumptions. Whether or not dialectal variation is fundamentally based on alternations and whether or not dialectometry models require alternations, the argument here is that the ability to distinguish between dialects (without topic-based features) is a rigorous evaluation of the quality of a model of dialects.

Finally, how does geographic variation as modeled here interact with register variation? We can think about this in two different ways: First, does register variation within these datasets present a confound by being structured geographically? In other words, if the corpus from Australia represents newspaper and magazine articles but the corpus from New Zealand represents discussion forums, then the ability to distinguish between the two is a confound. Given the size of the datasets, the consistent collection methodology, the cross-fold validation experiments, the large number of national varieties per language, and the comparison of web-based and Twitter data, however, this confound is not likely. Second, is register variation the same underlying phenomenon as regional variation? In other words, is the difference between New Zealand English and Australian English ultimately the same type of phenomenon as the structured difference between newspaper writing and discussion forums? This is an empirical question for future work that requires a dataset containing both register meta-data and spatial meta-data.

2.3. Modeling National Varieties

The experiments in this paper take a classification approach to dialectometry: given a one-hot encoding of construction frequencies (i.e., a bag-of-constructions), can we distinguish between different national varieties of a language? There are

[8] https://labbcat.canterbury.ac.nz/download/?jonathandunn/CxG_Data_FixedSize

two main advantages to this approach: First, the model can be evaluated using prediction accuracies on held-out testing data. This is important to ensure that the final model is meaningful. Second, a classification approach provides an implicit measure of the degree of syntactic separation between national varieties across the entire grammar (c.f., region similarities in section 3.3). A particular construction may be unique to a given variety, but this in itself is less meaningful if the varieties are otherwise the same. How deep or robust is the syntactic variation? How distinct are the national varieties? Dialectometry is about going beyond variation in individual syntactic features to measure the aggregate syntactic relationships between varieties.

The main set of experiments uses a Linear Support Vector Machine (Joachims, 1998) to classify varieties using CxG features. Parameters are tuned using separate development data[9]. Given the general robust performance of SVMs in the literature relative to other similar classifiers on latent variation tasks (Dunn et al., 2016), we forego a systematic evaluation of classifiers. For reproducibility against future work, all results are calculated on pre-specified training and testing sets. Given the large number of samples in each test set (**Table 2**) and the robustness in the cross-validation evaluation (**Table 4**) we are not concerned with over-fitting and given the high performance in general we are not concerned with under-fitting (**Table 3**). Under this evaluation regime, any classifier could be used; thus, it is not important to contrast a Linear SVM with other shallow classifiers such as Naive Bayes or Decision Trees in this context. The Linear SVM uses the training data to learn weights for each construction in the grammar for each regional variety; in the aggregate, the model builds a high-dimensional representation of each variety that maximizes the distance between them (i.e., so that varieties like American English and Nigerian English can be easily separated). The quality and generalizability of the models are evaluated using held-out testing data: how well can those same feature weights be used to predict which regional variety a new sample belongs to? Because it is possible here that the varieties could be distinguished in a low-dimensional space (i.e., being separated along only a few constructions), we use unmasking to evaluate the robustness of the models in section 3.2. This classification-based approach deals very well with the aggregation of features, including being able to ignore redundant or correlated features. On the other hand, this robust aggregation of syntactic features requires that we assume the spatial boundaries of each regional variety.

Moving to data preparation, the assumption is that a language sample from a web-site under the *.ca* domain originated from Canada. This approach to regionalization does not assume that whoever produced that language sample was born in Canada or represents a traditional Canadian dialect group; rather, the assumption is only that the sample represents someone in Canada who is producing language data; but the two are closely related (Cook and Brinton, 2017). This corresponds with the assumption that Twitter posts geo-referenced to particular coordinates represent language use in that place but do not necessarily represent language use by locals. Geo-referenced documents represent language use *in* a particular place. Unlike traditional dialect surveys, however, there is no assurance that individual authors are native speakers *from* that place. We have to assume that most language samples from a given country represent the native varieties of that country. For example, many non-local residents live in Australia; we only have to assume that most speakers observed in Australia are locals. On the one hand, this reflects the difference between corpus-based and survey-based research: we know less about the individuals who are represented in these datasets. On the other hand, this reflects increased mobility: the idea that a *local* individual is born, is raised, and finally dies all in the same location is no longer proto-typical.

TABLE 3 | F1 of classification of regional varieties by language and feature type (web corpus above and twitter corpus below).

CC	Function	CxG-1	CxG-2	Unigram	Bigram	Trigram	N. Regions
Arabic	0.88	0.90	1.00	1.00	1.00	0.96	4
English	0.65	0.80	0.96	1.00	0.98	0.87	14
French	0.61	0.78	0.96	1.00	0.98	0.90	13
German	0.84	0.89	0.96	1.00	0.98	0.86	8
Portuguese	0.89	0.98	0.99	1.00	1.00	0.97	4
Russian	0.41	0.79	0.95	1.00	0.95	0.80	19
Spanish	0.52	0.78	0.95	1.00	0.99	0.91	17
TW	**Function**	**CxG-1**	**CxG-2**	**Unigram**	**Bigram**	**Trigram**	**N. Regions**
Arabic	0.80	0.88	0.98	1.00	1.00	0.94	8
English	0.55	0.76	0.92	1.00	0.97	0.82	14
French	0.88	0.98	0.98	1.00	1.00	0.99	4
German	0.83	0.90	0.95	1.00	0.99	0.95	2
Portuguese	1.00	0.99	1.00	1.00	1.00	0.99	2
Russian	0.73	0.83	0.93	1.00	0.94	0.87	3
Spanish	0.51	0.82	0.94	1.00	0.99	0.92	17

In order to average out the influence of out-of-place samples, we use random aggregation to create samples of exactly 1,000 words in both corpora. For example, in the Twitter corpus this means that an average of 59 individual Tweets from a place are combined into a single sample. First, this has the effect of providing more constructions per sample, making the modeling task more approachable. Second and more importantly, individual out-of-place Tweets and web pages are reduced in importance because they are aggregated with other Tweets and web pages presumably produced by local speakers. If we think of non-locals as outliers, this approach aggregates outliers with non-outliers in order to reduce their influence. We leave for future work an evaluation of different approaches to this problem. The larger issue is the relationship between small but carefully curated corpora for which significant meta-data is available for each speaker and these large but noisy corpora which are known to contain out-of-place samples (i.e., tourists in Twitter data). One promising approach is to evaluate such noisy

[9] Development data allows experimental settings and parameters to be evaluated without over-fitting the training/testing data that is used for the main experiment.

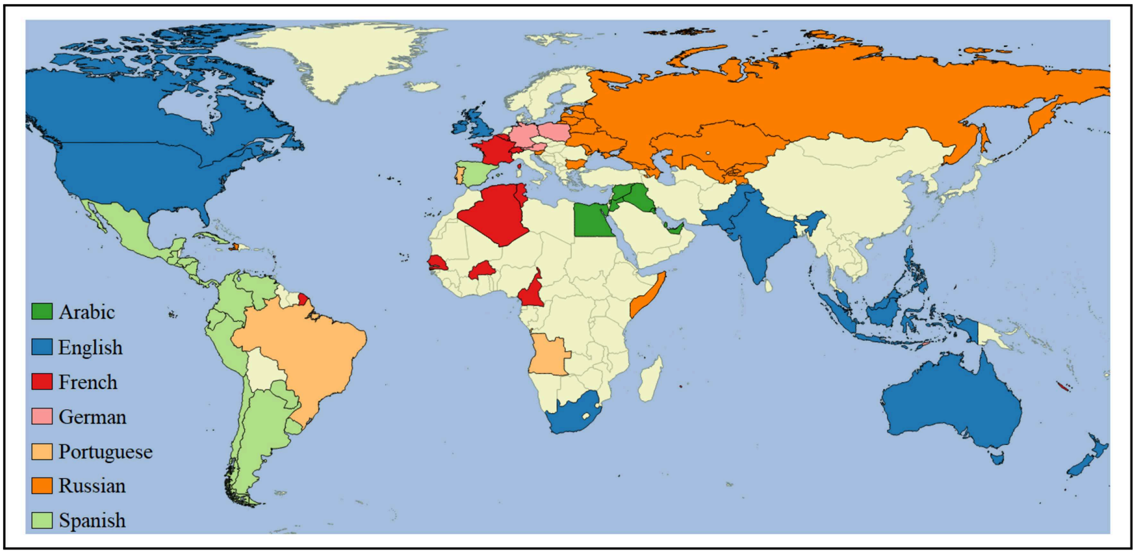


FIGURE 2 | Countries with national varieties for selected languages.

corpora based on how well they are able to predict demographic meta-data for the places they are intended to represent (Dunn and Adams, 2019). In this case, it has been shown that web-crawled and Twitter corpora are significantly correlated with population density (especially when controlling for GDP and general rates of internet usage) and that both datasets can be used to predict which languages are used in a country (as represented using census data). While there is much work to be done on this problem, the prediction of demographic meta-data provides a way to evaluate the degree to which large and noisy corpora reflect actual populations.

We take a simple threshold-based approach to the problem of selecting national varieties to include. For English and Spanish, any national variety that has at least 15 million words in both the Common Crawl and Twitter datasets is included. Given the large number of countries in **Table 2**, this higher threshold accounts for the fact that both English and Spanish are widely used in these datasets. Lower relative thresholds are used for the other languages, reflecting the more limited prevalence of these languages: the thresholds are made relative to the amount of data per language and are comparable to the English and Spanish threshold. For English and Spanish, the national varieties align across both datasets; thus, the experiments for these two languages are paired and we also consider similarity of models across registers. But for the other languages aligning the national varieties in this way removes too many from consideration; thus, there is no cross-domain evaluation for Arabic, French, German, Portuguese, or Russian.

The inventory of national varieties in **Table 2** is entirely data-driven and does not depend on distinctions like dialects vs. varieties, inner-circle vs. outer-circle, or native vs. non-native. Instead, the selection is empirical: any area with a large amount of observed English usage is assumed to represent a national variety of English. Since the regions here are based on national boundaries, we call these national varieties. We could just as easily call them national dialects or regional varieties. The global distribution of national varieties for each language is shown in **Figure 2**.

The datasets are formed into training, testing, and development sets as follows: First, 2k samples are used for development purposes regardless of the amount of data from a given variety. Depending on the size of each variety, at least 12k training and 2.5k testing samples are available. Because some varieties are represented by much larger corpora (i.e., Tweets from American English), a maximum of 25k training samples and 5k testing samples are allowed per variety per register. These datasets contain significantly more observations than have been used in previous work (Dunn, 2018a).

For each language, we compare six sets of features: First, syntactic representations using CxG-1 and CxG-2; Second, indirect syntactic representations using function words[10]; Third, unigrams and bigrams and trigrams of lexical items. Lexical unigrams represent mostly non-syntactic information while increasing the size of *n* begins to indirectly include information about transitions. The n-grams are representing using a hashing vectorizer with 30k dimensions (thus, these representations have no syntactic features present). This avoids biasing the selection of specific n-grams (i.e., with content more associated with dominant inner-circle varieties). But this also means that the lexical features themselves cannot be inspected.

3. RESULTS

This section reports the results of dialectometry experiments across seven languages. First, in section 3.1 we look at overall predictive accuracy using the F-Measure metric across feature

[10] For replicability, these are taken from https://github.com/stopwords-iso

sets and languages. The purpose of this analysis is to contextualize and then explore the interpretation of classification-based dialectometry. Second, in section 3.2 we examine the robustness of models across registers (using the web corpus and the Twitter corpus) and across rounds of feature pruning. The purpose of this analysis is to understand how meaningful these models are in the presence of possible confounds such as a reliance on a small number of highly predictive variants. These first two sections are important for validating a classification-based approach to syntactic variation. Third, in section 3.3 we analyze predictive accuracy and prediction errors across languages and develop representations of regional syntactic similarity. The purpose of this analysis is to use dialect classification to understand global syntactic variation in the aggregate. Fourth, in section 3.4 we examine measures of the uniqueness of different regional varieties and in section 3.5 we apply these models to evaluate empirical correlates for notions like *inner-circle* and *outer-circle*. These last two sections are important for understanding what dialect classification can tell us about global, whole-grammar syntactic variation once the approach itself has been validated.

3.1. Features, Model Size, and Predictive Accuracy

The overall prediction accuracy across languages is shown in **Table 3** (with the web corpus above and the Twitter corpus below). On the left-hand part of the table, the syntactic features are grouped: function words and the two CxG feature sets. On the right-hand part, the lexical features are grouped: lexical unigrams, bigrams, and trigrams. For reference, the number of regions for each variety is shown in the final column.

A classification-based approach has the goal of distinguishing between national varieties. We would expect, then, that the task of distinguishing between a small number of varieties is easier than distinguishing between a larger number of varieties. For example, there are only two varieties of German and Portuguese in the Twitter corpus. For Portuguese, all feature sets have F1s of 1.00 or 0.99; in other words, this is an easy task and there are many ways of doing it. This is also an indication that these varieties of Portuguese (here, from Brazil, BR, and from Portugal, PT) are quite distinct across all feature sets. On the other hand, even though German also has a small number of national varieties (here, from Germany, DE, and from Austria, AT), there is a wide variation in prediction accuracy, with function words (F1 = 0.83) and CxG-1 (F1 = 0.90) having markedly lower performance than other feature sets. The point is that model performance depends on both the number of national varieties included in the model (showing the importance of taking an empirical approach to the selection of varieties) as well as on the degree of difference between the varieties themselves. Portuguese as used in Brazil and Portugal is significantly more distinct than German as used in Germany and Austria. Digging deeper, however, we also notice that function words as features are more uneven across languages than other feature sets. For example, Arabic on Twitter has eight national varieties and function words achieve an F1 of 0.80; but for Russian on Twitter, with only three varieties, function words achieve a lower F1 of 0.73. This is an indication that, as indirect proxies for syntactic structure, the usefulness of function words for this task varies widely by language (at least, given the inventory of function words used here).

Regardless of the number of national varieties per language, lexical unigrams perform the best (F1 = 1.00). In other words, it is not difficult to disinguish between samples from New Zealand and Australia when given access to lexical items (*Christchurch* vs. *Brisbane*). While we know that syntactic models are capturing linguistic variation, however, the success of lexical models, as argued elsewhere (Dunn, 2019b), is partly a result of place-names, place-specific content, and place-specific entities. In other words, geo-referenced texts capture the human geography of particular places and this human geography information takes the form of specific lexical items. Previous work has focused on capturing precisely this type of content (Wing and Baldridge, 2014; Adams, 2015; Hulden et al., 2015; Lourentzou et al., 2017; Adams and McKenzie, 2018). The problem is that, without organizing the frequency of such lexical features according to concept (Zenner et al., 2012), these models may not represent linguistic variation[11]. For example, we know that as *n* increases n-grams represent increasing structural information (i.e., transitions between lexical items instead of lexical items in isolation). Here we see that, by the time *n* is raised to three, the predictive accuracy of CxG-2 always surpasses the predictive accuracy of trigrams (with the single exception of French on Twitter). The difference between CxG-2 and bigrams is much smaller than the distance between the various syntactic features. This is evidence that the advantage of unigrams over CxG-2 reflects the advantage of human geography content (i.e., lexical items in isolation) over linguistic variation (i.e., transitions between lexical items). In short, while some of the lexical variation is linguistic (*soda* vs. *pop*), a good deal of it is also based on human geography (*Chicago* vs. *Singapore*). The advantage of syntactic models in this context is that such non-linguistic variations do not introduce confounds: we know that these models represent regional varieties of each language.

Models on the web corpus (above) have higher predictive accuracy than models on the Twitter corpus (below). This is true except in cases, such as Portuguese, where there is a wide difference in the number of national varieties represented (for Portuguese, two vs. four). For reasons of data availability, only English and Spanish have strictly aligned varieties; in both of these languages, the syntactic features perform better on the web corpus than the Twitter corpus, although the gap is wider for English than for Spanish. This raises a question that is addressed in the next section: are models of syntactic variation consistent across these registers? In other words, do the web-based and Twitter-based models make the same types of errors?

The web corpus also provides more varieties per language (with Arabic as the sole exception, which is better represented on Twitter). In many cases this difference is significant: there are 19 varieties of Russian on the web, but only three on Twitter.

[11] This is a simplification, of course, but the underlying point is that it is difficult to distinguish linguistic lexical variation from human geography-based and topical lexical variation without relying on the idea of conceptual alternations.

In this case, there are competing Russian-language social media platforms (i.e., www.vk.com) that are not included in this study. In other words, outside of English and Spanish, which are aligned across datasets, the Twitter data is less comprehensive.

What does the F-Measure tell us about models of syntactic variation? First, the measure is a combination of precision and recall that reflects the predictive accuracy while taking potentially imbalanced classes into account: how many held-out samples can be correctly assigned to their actual region-of-origin? On the one hand, this is a more rigorous evaluation than simply finding a significant difference in a syntactic feature across varieties within a single-fold experimental design: not only is there a difference in the usage of a specific feature, but we can use the features in the aggregate to characterize the difference between national varieties. On the other hand, it is possible that a classifier is over-fitting the training data so that the final model inflates the difference between varieties. For example, let's assume that there is a construction that is used somewhat frequently in Pakistan English but is never used in other varieties. In this case, the classifier could achieve a very high prediction accuracy while only a single construction is actually in variation. Before we interpret these models further, the next section evaluates whether this sort of confound is taking place.

3.2. Model Robustness Across Features and Registers

If a classification model depends on a small number of highly predictive features, thus creating a confound for dialectometry, the predictive accuracy of that model will fall abruptly as such features are removed (Koppel et al., 2007). Within authorship verification, *unmasking* is used to evaluate the robustness of a text classifier: First, a linear classifier is used to separate documents; here, a Linear SVM is used to classify national varieties of a language. Second, for each round of classification, the features that are most predictive are removed: here, the highest positive and negative features for each national variety are pruned from the model. Third, the classifier is retrained without these features and the change in predictive accuracy is measured: here, unmasking is run for 100 iterations using the CxG-2 grammar as features, as shown in **Figure 3** (with the web-based model above and the Twitter-based model below). For example, this removes 28 constructions from the model of English each iteration (two for each national dialect), for a total of approximately 2,800 features removed. The figures show the F-Measure for each iteration. On the left-hand side, this represents the performance of the models with all features are present; on the right-hand side, this represents the performance of the models after many features have been removed. This provides a measure of the degree to which these models are subject to a few highly predictive features.

First, we notice that models with a higher starting predictive accuracy (e.g., Arabic and Portuguese in the web-based model and Portuguese and French in the Twitter-based model) tend to maintain their accuracy across the experiment. Even after 100 rounds of pruning, Arabic and Portuguese (CC) remain above 0.95 with CxG-2 features[12]. Similarly, French and Portuguese remain above 0.95 after 100 rounds of pruning (TW). This indicates that a high performing dialect classification model is based on a broad and distributed set of features. But this is not always the case: for example, Arabic (TW) starts out with the same performance as French but over the course of the experiment declines to a performance that is 10% lower than French. This is an indication that this Twitter-based model of Arabic is less robust than its counter-part model of French (although keep in mind that the French model has only 4 varieties and the Arabic model has 8).

Second, although Spanish and Russian have a starting accuracy that is comparable to other languages, with F1s of 0.95 for both languages on the web corpus, their accuracy falls much more quickly. Spanish and Russian decrease by around 20% by the end of the experiment while English and French decrease by only 10% in total. On the Twitter corpus, Spanish and Russian again pattern together, this time with a 15% reduction. But here the English model has a somewhat steeper decline. In most cases, however, the starting accuracy of a model is related to its rate of decline: more accurate models are also more robust to feature pruning. The purpose of this evaluation is to show that a classification approach to dialectometry is not subject to the confound of a small number of highly predictive features.

The next question is about the similarity of national varieties as represented in the web corpus vs. the Twitter corpus. Is there a consistent representation of variation or are the models ultimately register-specific? For this analysis we focus on English and Spanish as the two languages that are aligned by national varieties across both datasets. We focus on an analysis of errors: First, two national varieties that are more often confused by the classifier are more similar according to the model. Thus, we represent the similarity of regions using the total of all errors between two varieties. For example, if UK English is predicted to be New Zealand English 50 times and New Zealand English is predicted to be UK English 25 times, there are 75 total errors between these varieties. More errors reflects more similar varieties[13].

The question is whether the web corpus and Twitter both provide the same patterns of similarity. **Figure 4** shows the relative errors between varieties for both datasets (with English above and Spanish below): the web (blue) occupies the left-hand side of each bar and Twitter (red) occupies the right-hand side. If both colors are the same size, we see the same proportion of errors for a given pair across both datasets. This figure also shows the most similar varieties, with the varieties having the highest total errors occupying the bottom of each. For example, the most similar varieties of English on Twitter are American (US) and Canadian English (CA). The most similar varieties on the web corpus, however, are New Zealand (NZ) and South African English (ZA)[14]. The Pearson correlation between errors, paired across datasets by varieties, is highly significant for English

[12] Here and below we focus on CxG-2 as the highest performing syntactic model.

[13] Country abbreviations are given in Appendix A (**Supplementary Material**).

[14] The ISO country codes are used in all figures and tables; these are shown by common name in the first Appendix in **Supplementary Material**.

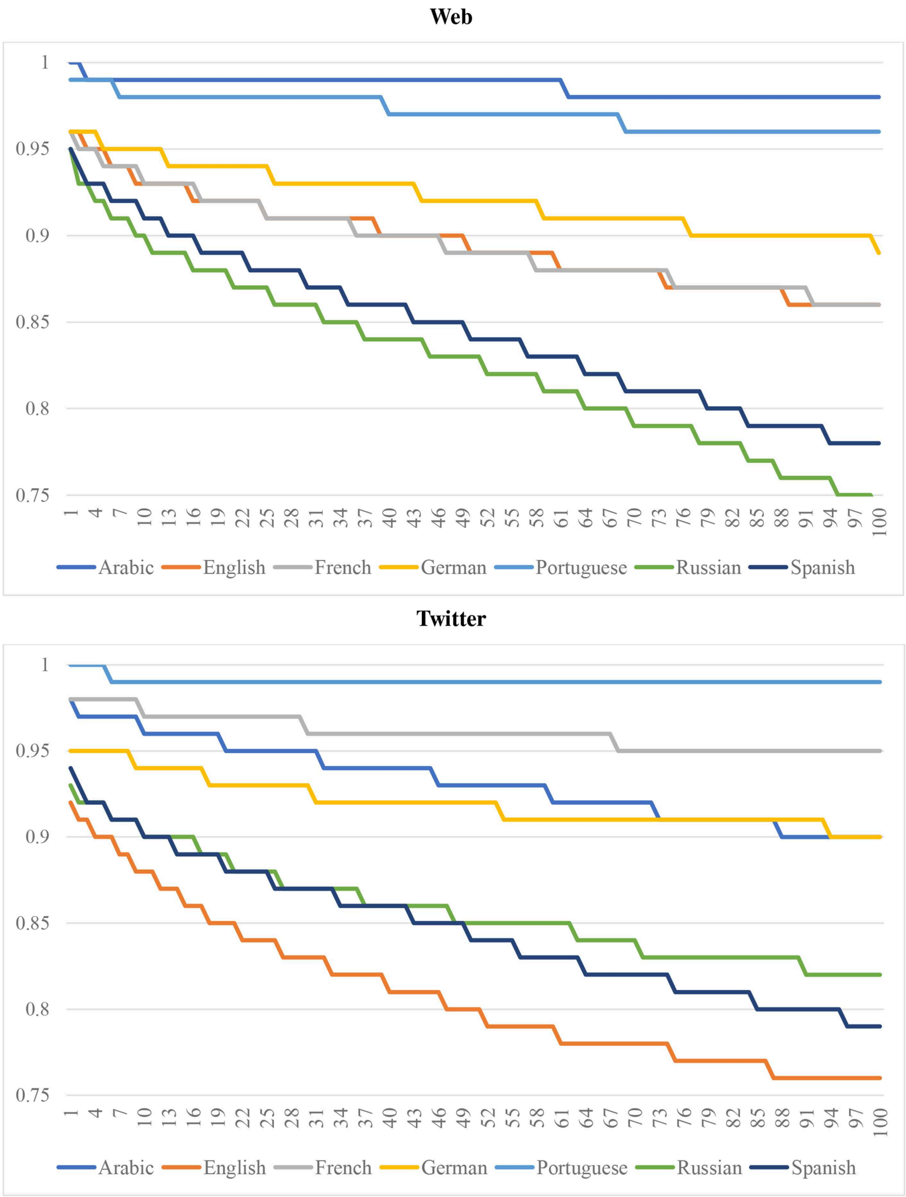


FIGURE 3 | Model robustness by language using unmasking for 100 iterations with CxG-2 features (web models above and twitter models below).

at 0.494 (note that this involves the number of errors but does not require that the errors themselves match up across registers). At the same time, there remain meaningful differences between the datasets. For example, Nigeria (NG) and Portugal (PT) have many errors in the Twitter model but very few in the web model. On the other hand, New Zealand (NZ) and South Africa (ZA) have a large number of errors in the web model but few in the Twitter model. This is an indication that the models are somewhat different across registers.

The errors for Spanish, in the bottom portion of **Figure 4**, also are significantly correlated across registers, although the Pearson correlation is somewhat lower (0.384). For example, both corpora have significant errors between Argentina (AR) and Uruguay (UY), although Twitter has a much higher error rate. But errors between Costa Rica (CR) and Uruguay (UY) and between Argentina (AR) and Costa Rica (CR) are only found on Twitter. Errors between Honduras (HN) and Nicaragua (NI), on the other hand, are only found in the web model. The point is that the two registers are associated in their error rates for both English and Spanish (the only languages with regional varieties aligned across both datasets).

The high accuracy of these models could suggest that the models are over-fitting the test set, even with a relatively large number of samples in the test set. Thus, in **Table 4**, we compare the weighted F1 scores on the test set with a 10-fold cross-validation evaluation that includes the training and testing data

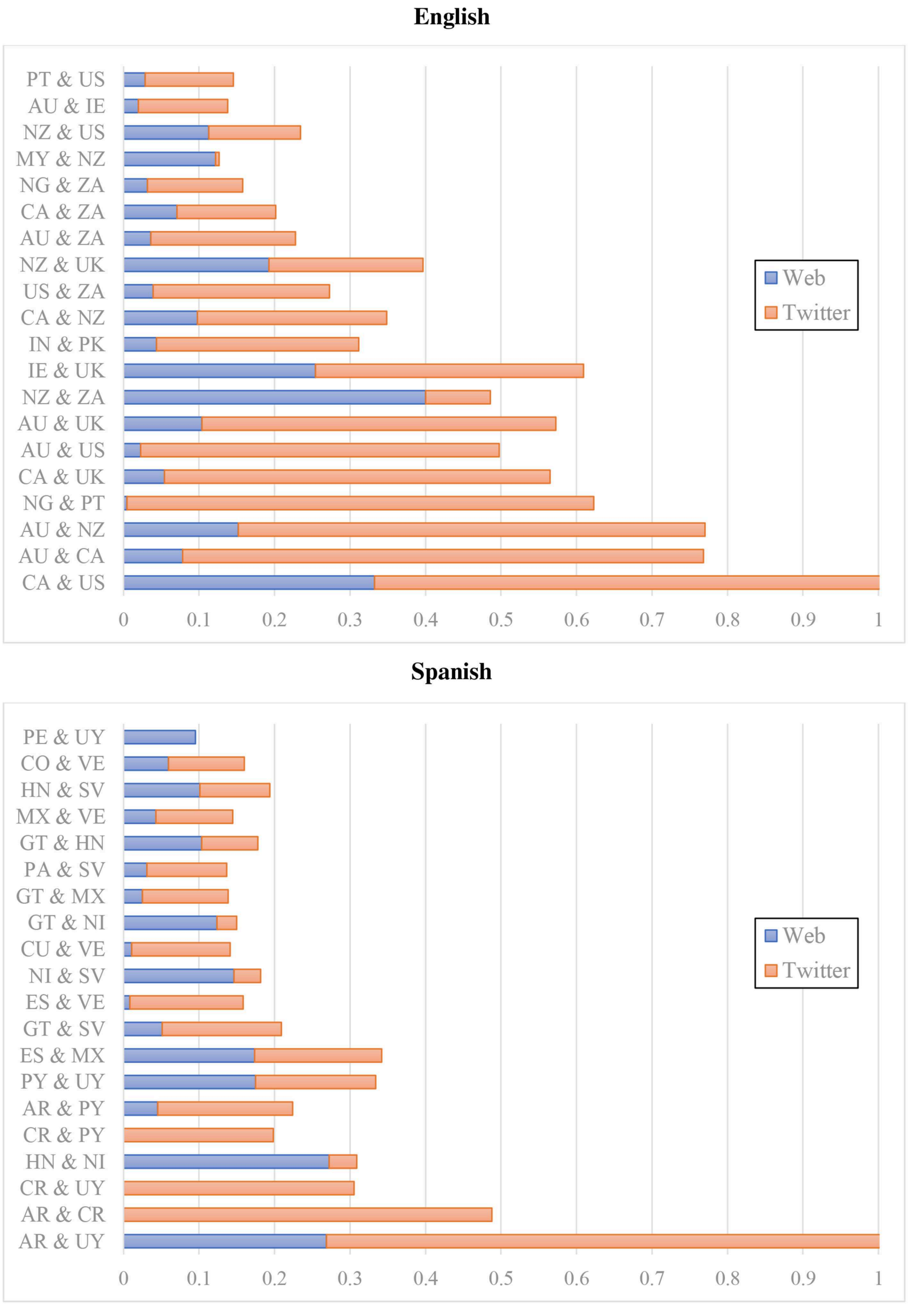


FIGURE 4 | Classification errors by percent of dataset for web and twitter corpora using CxG-2 features (English errors above and Spanish errors Below).

together. The table shows the maximum and minimum values across folds. There are only three cases in which the minimum fold F1 is lower than the reported test set metrics: Russian (web data), Arabic (Twitter data), and Portuguese (Twitter data). In each case the difference is small and in each case the average fold F1 is the same as the F1 from the test set alone. This evidence shows that the models are not over-fitting the test set and that this reflects a robust classification accuracy.

This section has approached two important questions: First, is a classification model dependent on a small number of highly predictive features? Second, does a classification model produce the same type of errors across both web corpora and Twitter corpora? In both cases some languages (like English) are more robust across feature pruning and more stable across registers than others (like Spanish). This is the case even though the F-Measure (reflecting predictive accuracy alone) is similar for both languages: 0.96 vs. 0.95 for the web model and 0.92 vs. 0.94 for the Twitter model. These alternate evaluations, then, are important for revealing further properties of these classification models. The predictive accuracy for both languages is high across

TABLE 4 | Comparison of train/test and cross-validation results by weighted F1 for CxG-2.

	CC				TW		
	Train-Test	CV-Max	CV-Min		Train-Test	CV-Max	CV-Min
Arabic	1.00	1.00	1.00	Arabic	0.98	0.98	**0.97**
English	0.96	0.96	0.96	English	0.92	0.92	0.92
French	0.96	0.96	0.96	French	0.98	0.98	0.99
German	0.96	0.96	0.96	German	0.95	0.96	0.95
Portuguese	0.99	0.99	0.99	Portuguese	1.00	1.00	**0.99**
Russian	0.95	0.95	**0.94**	Russian	0.93	0.95	0.93
Spanish	0.95	0.95	0.95	Spanish	0.94	0.94	0.94

Bold values indicate CV results lower than results on the test set.

TABLE 5 | Classification performance for English regions, web, and twitter corpora, CxG-2 features.

	Prec (CC)	Recall (CC)	F1 (CC)		Prec (TW)	Recall (TW)	F1 (TW)
AU	0.97	0.96	0.97	AU	0.82	0.83	0.83
CA	0.94	0.94	0.94	CA	0.84	0.79	0.81
IE	0.97	0.97	0.97	IE	0.95	0.95	0.95
NZ	0.91	0.92	0.91	NZ	0.92	0.90	0.91
UK	0.95	0.95	0.95	UK	0.87	0.90	0.89
US	0.93	0.95	0.94	US	0.85	0.89	0.87
ZA	0.94	0.96	0.95	ZA	0.92	0.94	0.93
IN	0.97	0.98	0.97	IN	0.97	0.97	0.97
MY	0.96	0.96	0.96	MY	0.99	0.99	0.99
NG	0.98	0.98	0.98	NG	0.94	0.95	0.94
PH	0.98	0.97	0.98	PH	0.98	0.98	0.98
PK	1.00	0.99	0.99	PK	0.98	0.98	0.98
CH	0.97	0.94	0.96	CH	0.98	0.97	0.97
PT	0.99	0.98	0.98	PT	0.93	0.90	0.92
AVG	0.96	0.96	0.96	AVG	0.92	0.92	0.92

both registers and the regional varieties which are confused is significantly correlated across both registers.

3.3. Regional Accuracy and Similarity

While the previous sections have evaluated classification-based models externally (prediction accuracy by feature type, robustness across feature pruning, error similarity across registers), this section and the next focus on internal properties of the models: what are the relationships between national varieties for each language? Which regions perform best within a model? In this section we examine the F-Measure of individual national varieties and the similarity between varieties using cosine similarity between feature weights. Because the Twitter dataset has fewer varieties for most languages, we focus on similarity within the web models alone and only for languages with a large inventory of varieties (i.e., only for English, French, and Spanish).

We start with English in **Table 5**. The left-hand side shows Precision, Recall, and F-Measure scores for the web corpus and the right-hand side for the Twitter corpus, both using the CxG-2 feature set. The higher the scores for each national dialect, the more distinct that variety is from the others in syntactic terms. New Zealand English (NZ) has the lowest F1 (0.91) for the web corpus. While the score of NZ English is the same for the Twitter model (0.91), it is no longer the lowest scoring variety: this is now Canadian English (CA) at 0.81. In fact, the lowest performing varieties for the Twitter model are all *inner-circle* varieties: Australia (AU), Canada (CA), United Kingdom (UK), and the United States (US). This phenomenon is explored further in the next section: why are more dominant varieties more difficult to model? Is this consistent across languages? For now we note only that all of the countries included in the model are expected, with perhaps the exception of Portugal (PT) and Switzerland (CH). While previous work made an explicit distinction between inner-circle and outer-circle varieties (Dunn, 2018a), here we leave this type of categorization as an empirical question.

We can compare national varieties within a model by comparing their respective feature weights: which regions have the most similar syntactic profiles? We use cosine distance to

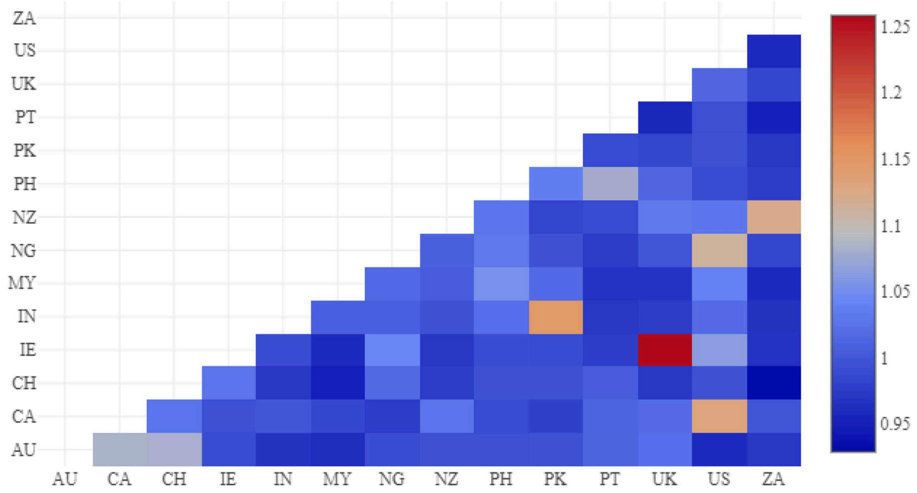


FIGURE 5 | Region similarity by cosine between feature weights, English CxG-2.

measure the similarity between feature weights and then use a heat map, as in **Figure 5**, to visualize the similarities. Cells with a higher value (more red) indicate a pair of varieties which the model is trying hard to separate (thus, a more similar pair). For example, the most similar pair is UK English (UK) and Irish English (IE); this is expected given that Northern Ireland is part of the UK. The next four pairs also are expected: Indian (IN) and Pakistan English (PK), American (US) and Canadian English (CA), New Zealand (NZ) and South African English (ZA), American (US) and Nigerian English (NG). While the final pair is less transparent, it is important that the model picks out these pairs of related varieties without any pre-knowledge. On the other hand, dark blue values indicate that the model is not concerned with separating the pair (because they are not very similar): for example, South African English (ZA) and Swiss English (CH).

French varieties are shown in **Table 6**, with again a much larger inventory for the web model than for the Twitter model. As with English, the lowest performing varieties in terms of prediction accuracy are the most dominant inner-circle varieties: France (FR), Belgium (BE), and Switzerland (CH). One possible reason is that there is more internal variation in France than in, for example, Cameroon (CM). Another possible reason is that these inner-circle varieties have influenced the outer-circle varieties, so that they are harder to distinguish from the colonial varieties. The regions in the web model are expected given French colonial history: European varieties (France, Switzerland, Belgium, Luxembourg), African varieties (Burkina Faso, Cameroon, Senegal), North African varieties (Grenada, Algeria, Tunisia), Pacific varieties (New Caledonian, French Polynesia), and unconnected island varieties with current or past French governance (Réunion, Grenada). All have a history of French usage.

Following the same methodology for English, region similarity is shown in **Figure 6**. The closest varieties are from Réunion and French Polynesia, from Senegal and Burkina Faso, and from France and Belgium. This again shows that the model not only distinguishes between varieties but can also situate the varieties in relationship to one another.

Next, regional accuracies for Spanish are shown in **Table 7**; these are aligned by country with the exception of Peru (PE) which is missing from the Twitter dataset. There is a single European variety (Spain), South American varieties (Argentina, Chile, Colombia, Ecuador, Peru, Paraguay, Uruguay, Venezuela), Central American varieties (Costa Rica, Guatemala, Honduras, Nicaragua, Panama, El Salvador), as well as Cuban and Mexican varieties. The alignment across datasets helps to ensure that only expected varieties occur; as discussed above, there is in fact a significant correlation between the errors produced on the two datasets.

The similarity between Spanish regions is shown in **Figure 6** (below French). The most similar varieties are from Costa Rica and Chile, from Spain and Chile, and from Venezuela and Colombia. The least similar are from Argentina and Chile and from Peru and Venezuala.

Russian varieties are shown in **Table 8**, encompassing much of Eastern Europe and Central Asia. As mentioned before, the Twitter dataset is missing a number of important varieties, most likely because of the influence of other social media platforms. There are two noisy regions, SO and PW, present in the web corpus[15]. Beyond this, the countries represented are all expected: in addition to Russia (RU), there are varieties from Central Asia (Azerbaijan, Georgia, Kyrgyzstan, Tajikistan, Uzbekistan), Southeast Europe (Bulgaria, Moldova), and Eastern Europe (Belarus, Lithuania, Slovenia, Ukraine). There are also varieties that reflect expanding-circle varieties of Russian (Ecuador, Haiti). Given the lack of alignment between the datasets, it is difficult to evaluate whether or not these expanding-circle varieties are

[15] One approach that could remove the few noisy regions that show up in Russian and, later, in German is to use population-based sampling to reduce the amount of data per country before selecting regional varieties.

TABLE 6 | Classification performance for French regions, web, and twitter corpora, CxG-2 features.

	Prec (CC)	Recall (CC)	F1 (CC)		Prec (TW)	Recall (TW)	F1 (TW)
BE	0.94	0.86	0.90	BE	0.97	0.94	0.96
BF	0.98	0.98	0.98	BF	–	–	–
CH	0.92	0.93	0.93	CH	–	–	–
CM	1.00	1.00	1.00	CM	–	–	–
DZ	0.99	0.99	0.99	DZ	–	–	–
FR	0.92	0.95	0.93	FR	0.97	0.98	0.98
GD	0.94	0.92	0.93	GD	–	–	–
HT	–	–	–	HT	1.00	1.00	1.00
LU	0.97	0.96	0.96	LU	1.00	1.00	1.00
NC	0.96	0.95	0.95	NC	–	–	–
PF	0.97	0.97	0.97	PF	–	–	–
RE	0.94	0.95	0.95	RE	–	–	–
SN	0.98	0.98	0.98	SN	–	–	–
TN	0.98	0.97	0.98	TN	–	–	–
AVG	0.96	0.96	0.96	AVG	0.98	0.98	0.98

robust. This reflects another limitation of an entirely data-driven approach: when is the use of Russian in a country a stable dialect and when is it a non-native variety that reflects short-term military or economic connections? The capacity of this syntactic model to predict both suggests that, in empirical terms, the distinction is not important. It could be the case, however, that some varieties are more robust than others to feature pruning. For reasons of space, similarities between Russian varieties are not shown.

Because they have fewer national varieties each, we end with Arabic, German, and Portuguese together (this table is shown in Appendix 2 (**Supplementary Material**)). Starting with Arabic, the regional comparison is made difficult by the little overlap between the two datasets: only data from Syria is consistent across registers. Focusing on the Twitter model, then, we note that it does contain examples of several traditional dialect groups: Algerian (DZ) represents the Maghrebi group, Egypt (EG) represents the Egyptian group, Iraq (IQ) and Syria (SY) represent the Mesopotamian group, Jordan (JO) and Palestine (PS) represent the Levantine group, and Kuwait (KW) represents the Arabian group. In addition, there is a Russian (RU) dialect of Arabic, reflecting an emerging outer-circle variety. Given the sparsity of regions shared across the two datasets, we do not explore further the relationships between varieties. The point here is to observe that the models on both datasets maintain a high accuracy across regions and that the available countries do represent many traditional dialect groups.

For German, Twitter provides only a few inner-circle varieties. Here we see, again, that the most central or proto-typical dialect (Germany, DE) has the lowest overall performance while the highest performance is found in less-central varieties. While other languages have national varieties representing countries that we expect to see, the German web corpus contains three regions that are almost certainly noise: the PW (Palau), SO (Somalia), and TL (East Timor) domains are most likely not used for regional web pages but rather for other purposes. No other language has this sort of interference by non-geographic uses of domain names (except that Russian also picks up data from *.so* and *.pw*). Most likely this results from having a frequency threshold that is too low. Because a classifier attempts to distinguish between all classes, the inclusion of noisy classes like this may reduce performance but will never improve performance. Thus, we leave this model as-is in order to exemplify the sorts of problems that an entirely data-driven methodology can create. Ignoring these varieties, however, the web-based model does provide a well-performing model of Austria (AU), Switzerland (CH), Germany (DE), Luxembourg (LU), and Poland (PL).

For Portuguese, again the Twitter model only covers major varieties: Brazil and Portugal. The web corpus, unlike German, does not show any noisy regions but it does include two expected African varieties: Angola (AO) and Cabo Verde (CV). While the model performs well, we will not delve more deeply into the region-specific results.

The purpose of this section has been to examine the prediction accuracies across national varieties alongside the similarity between varieties. With the exception of some noisy regions for German and Russian, these results show that the model both is able to make accurate predictions about syntactic variation as well as to make reasonable representations of the aggregate similarity between national varieties.

3.4. Empirical Measures of Region Uniqueness

We have seen in the sections above that outer-circle or expanding-circle varieties often have higher predictive accuracies even though they are less proto-typical and less dominant. For example, these sorts of varieties have been shown to have lower feature densities for these CxG grammars (Dunn, 2019b), which indicates that the grammars are missing certain unique constructions. Regardless, these varieties remain unique in that they are easier to distinguish from more central varieties.

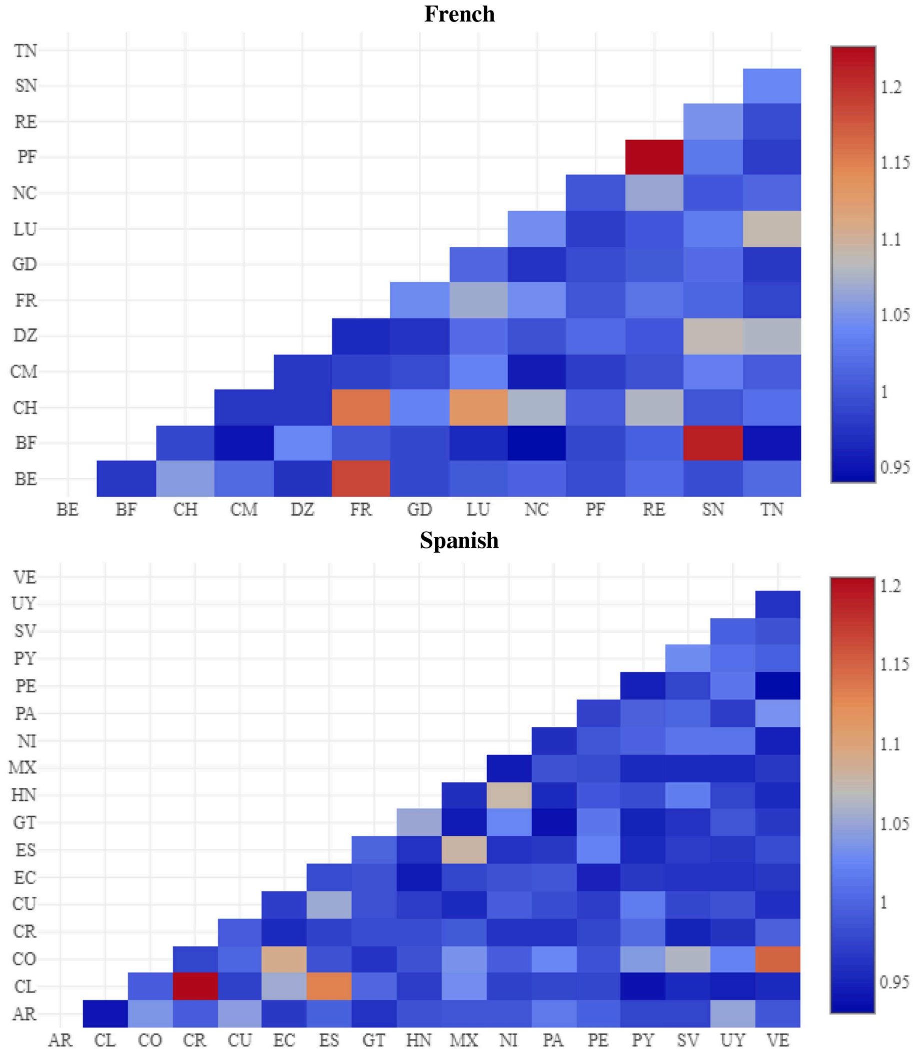


FIGURE 6 | Region similarity by cosine between feature weights, French (above) and Spanish (below) CxG-2.

For example, the English Twitter models show the main inner-circle varieties as having the lowest F1 scores: Australia (0.83), Canada (0.81), United States (0.87), and the United Kingdom (0.89). This phenomenon is not limited to English, however. In the French web model, again the inner-circle (i.e., European) varieties have the lowest F1 scores: Belgium (0.90), Switzerland (0.93), and France (0.93). The other languages do not present examples as clear as this; for example, Arabic and German and Portuguese do not contain enough varieties to make such a comparison meaningful. Russian and Spanish are characterized by a large number of varieties that are contiguous in relatively dense regions, thus showing a less striking colonial pattern. Why is it that, in cases of non-contiguous dialect areas, the inner-circle varieties have the lowest prediction accuracy?

In qualitative terms, there are several possible explanations. First, it could be the case that these inner-circle varieties have strongly influenced the other varieties so that parts of their syntactic profiles are replicated within the other varieties. Second, it could be that there is an immigration pipeline from outer-circle to inner-circle countries, so that the samples of UK English, for example, also contain speakers of Nigerian English. Third, it could be the case that media and communications are centered around inner-circle markets so that outer-circle varieties are influenced by one or another center of power. Additional factors could include the strength of standardization across languages, the number of L1 vs. L2 speakers that are represented for each language, and the average level of education for each country. None of these possibilities can be distinguished in empirical terms within the current study.

We have shown above, however, that this approach to dialectometry can (i) make accurate predictions about variety membership and (ii) can create reasonable representations of aggregate syntactic similarity between regions. In this section we formulate an approach to identifying, in purely synchronic terms,

TABLE 7 | Classification performance for Spanish regions, web, and twitter corpora, CxG-2 features.

	Prec (CC)	Recall (CC)	F1 (CC)		Prec (TW)	Recall (TW)	F1 (TW)
AR	0.94	0.94	0.94	AR	0.85	0.90	0.87
CL	0.99	0.98	0.98	CL	0.97	0.98	0.97
CO	0.95	0.94	0.95	CO	0.95	0.93	0.94
CR	1.00	1.00	1.00	CR	0.91	0.87	0.89
CU	0.96	0.97	0.97	CU	0.98	0.97	0.98
EC	0.96	0.96	0.96	EC	0.98	0.98	0.98
ES	0.94	0.95	0.94	ES	0.94	0.96	0.95
GT	0.96	0.96	0.96	GT	0.94	0.95	0.95
HN	0.93	0.94	0.94	HN	0.94	0.92	0.93
MX	0.94	0.93	0.93	MX	0.92	0.93	0.93
NI	0.92	0.86	0.89	NI	0.98	0.98	0.98
PA	0.98	0.98	0.98	PA	0.95	0.95	0.95
PE	0.94	0.92	0.93	PE	–	–	–
PY	0.94	0.96	0.95	PY	0.93	0.94	0.93
SV	0.95	0.94	0.95	SV	0.93	0.94	0.93
UY	0.91	0.93	0.92	UY	0.88	0.85	0.86
VE	0.97	0.98	0.98	VE	0.94	0.93	0.93
AVG	0.95	0.95	0.95	AVG	0.94	0.94	0.94

TABLE 8 | Classification performance for Russian regions, web, and twitter corpora, CxG-2 features.

	Prec (CC)	Recall (CC)	F1 (CC)		Prec (TW)	Recall (TW)	F1 (TW)
AZ	0.94	0.94	0.94	AZ	–	–	–
BG	1.00	1.00	1.00	BG	–	–	–
BY	0.98	0.95	0.97	BY	0.91	0.85	0.88
EC	0.96	0.98	0.97	EC	–	–	–
EE	0.86	0.89	0.87	EE	–	–	–
GE	0.95	0.95	0.95	GE	–	–	–
HT	0.99	0.99	0.99	HT	–	–	–
KG	0.99	0.99	0.99	KG	–	–	–
KZ	0.96	0.93	0.94	KZ	–	–	–
LT	0.94	0.93	0.94	LT	–	–	–
LV	0.92	0.91	0.91	LV	–	–	–
MD	0.98	0.97	0.97	MD	–	–	–
RU	0.90	0.90	0.90	RU	0.93	0.96	0.94
SI	1.00	1.00	1.00	SI	–	–	–
TJ	0.95	0.97	0.96	TJ	–	–	–
UA	0.93	0.94	0.94	UA	0.98	0.96	0.97
UZ	0.92	0.92	0.92	UZ	–	–	–
AVG	0.95	0.95	0.95	AVG	0.94	0.94	0.93

which varieties within a model represent central inner-circle countries that are the sources of influence for other outer-circle countries. The observations about prediction accuracy depend on the evaluation of the model, but we want this measure of uniqueness to depend on the model of variation itself.

The feature weights represent the positive and negative importance of each syntactic feature for each national variety. We used cosine similarities between feature weights above to find the most similar regions. Here we are interested in the overall uniqueness of a particular dialect: which varieties are in general not similar to any other varieties? We calculate this by summing the Spearman correlations between each variety and all other varieties. For example, if UK English has similar ranks of features as Irish and New Zealand English, then this will produce a high value. But if Swiss English generally has low relationships between feature ranks with other varieties, then this

TABLE 9 | Variety uniqueness by language using spearman correlation, web CxG-2 model.

English			French			Russian			Spanish		
1	US	-0.46	1	FR	-0.49	1	TJ	0.04	1	ES	-0.24
2	UK	-0.25	2	RE	-0.36	2	EE	0.15	2	PY	-0.05
3	CA	-0.22	3	CH	-0.32	3	SI	0.17	3	UY	-0.04
4	NZ	-0.18	4	LU	-0.26	4	LT	0.23	4	AR	-0.02
5	AU	-0.16	5	PF	-0.18	5	EC	0.23	5	CO	0.00
6	IE	-0.14	6	SN	-0.15	6	KZ	0.23	6	CL	0.03
7	PH	-0.06	7	BE	-0.10	7	UA	0.23	7	HN	0.04
8	MY	-0.05	8	NC	-0.08	8	LV	0.26	8	CU	0.06
9	IN	-0.01	9	BF	-0.03	9	GE	0.32	9	MX	0.10
10	NG	0.02	10	GD	-0.02	10	HT	0.35	10	NI	0.12
11	CH	0.05	11	TN	0.05	11	KG	0.35	11	GT	0.13
12	ZA	0.06	12	DZ	0.07	12	UZ	0.36	12	SV	0.15
13	PT	0.13	13	CM	0.25	13	AZ	0.36	13	CR	0.18
14	PK	0.14	–	–	–	14	BY	0.47	14	VE	0.19
–	–	–	–	–	–	15	RU	0.56	15	EC	0.23
–	–	–	–	–	–	16	MD	0.67	16	PE	0.25
–	–	–	–	–	–	17	BG	0.84	17	PA	0.32

will produce a low value. These uniqueness values are shown in **Table 9** for each of the languages with a large number of varieties, calculated using CxG-2 web-based models. Spearman correlations are preferred here instead of Pearson correlations because this reduces the impact of the distance between varieties (which the classifier is trying to maximize).

The uniqueness of each region reflects, at least for non-contiguous languages like English and French, the degree to which a variety belongs in the inner-circle. For example, the top three countries for English are the United States, the UK, and Canada; for French they are France, Réunion (the only French overseas department in the model), and Switzerland. In both cases the uniqueness of varieties with this measure reflects the same scale that categorizations like inner and outer circle are attempting to create. The most unique variety of Spanish is the only non-contiguous variety (from Spain). The interpretation of the rest of the regions on this scale is made more difficult because they are of course densely situated. Notice, also, that while English and French have a scale with higher uniqueness (with starting values of -0.46 and -0.49), both Spanish and Russian have a scale with higher similarity (with ending values of 0.84 and 0.32). Russian has no negative values at all, for example. The most unique varieties of Russian are from Tajikistan, Estonia, and Slovenia. Rather than being inner-circle, as in French and English, these are more peripheral varieties. While this uniqueness measure still reflects an important property of the relationships between varieties, then, its interpretation is complicated by the different behavior of languages with contiguous or non-contiguous varieties.

The purpose of this section has been to show that the feature weights from the model can also be used to create a general measure of variety uniqueness which reflects an important property of the status of varieties. While qualitative work creates categories like inner-circle or outer-circle, this produces a scale that represents similar intuitions. The difference is that the notion of inner-circle depends on historical and social information about variety areas, with little linguistic analysis, while this scale is entirely linguistic with no historical information whatsoever.

3.5. Empirical Evidence for World Englishes

How can we connect data-driven approaches to syntactic variation with qualitative assessments within sociolinguistics? In this section we compare the model of English variation in this paper with traditional classifications from the World Englishes paradigm into inner-circle, outer-circle, and expanding-circle varieties.

First we look at classification accuracy (c.f., **Table 5**). We expect that inner-circle varieties will be more closely clustered together as they are more closely related and are used in mainly monolingual contexts. There is a significant difference between inner-circle and outer-circle performance in both datasets using a two-tailed t-test ($p = 0.0183$ for CC and $p = 0.004$ for TW). Upon inspection we see that the outer-circle varieties have higher accuracies, in part because they are more unique.

Second, we look at the degree of fit between the grammar and each regional variety using the relative average frequency: how often do constructions in the grammar occur in each variety? In other words, because the grammar is learned on a different dataset which is likely skewed toward inner-circle varieties, we would expect that the grammar itself would better describe these varieties. A higher average frequency means a better default description (i.e., because the samples are all the same length and so should contain approximately the same number of constructions per sample). We again divide the varieties into inner-circle and outer-circle and test the significance of this difference using a two-tailed t-test: the result is significant ($p = 0.0011$ for CC and $p = 0.0004$ for TW). In this case, inspection shows that the inner-circle varieties have higher frequencies than the outer-circle varieties.

Third, we look at uniqueness values as calculated in **Table 9**. First, we see that there is a clear separation between inner-circle and outer-circle varieties, with the exception of South African English. But is the difference significant? Again using a two-tailed t-test there is a significant difference, although to a lesser degree $p = 0.024$ for CC).

In all three cases, there is a significant difference between attributes of inner-circle and outer-circle varieties: the prototypical inner-circle varieties are better described by the grammar but less distinguishable in terms of classification accuracy and in terms of aggregate similarities. There is a consistent and significant distinction, even when the model of varieties of English makes no geographic or sociohistorical assumptions.

4. DISCUSSION

This paper has shown (i) that whole-grammar dialectometry and data-driven language mapping can be brought together to produce models capable of predicting the membership of held-out samples with a high degree of accuracy. In addition, we have shown (ii) that these models do not depend on only a small number of highly predictive variants, (iii) that there is a strong association between classification errors across registers in those languages that are paired across both datasets, (iv) that the models can be used to create reasonable representations of the aggregate similarity between varieties, and (v) that measures of uniqueness based on these models provide an empirical approximation of categorical notions like inner-circle vs. outer-circle varieties. Taken together, these results show that a computational approach to dialectology can overcome the limitations of traditional small-scale methods. The discussion in this section focuses on two questions: First, how do these computational models of dialect relate to previous qualitative understandings of dialect? Second, what does the increased scale and scope of these models mean for interactions between sociolinguistics and computational linguistics?

4.1. Categorizing Varieties

At its core, the goal of computational dialectology is to provide precise global-scale models of regional linguistic variation that are both replicable and falsifiable. In other words, these models are *descriptions* of how linguistic structure (specifically, syntax as represented by CxG) varies across national varieties. But we also want to *explain* linguistic variation in historical or social terms: what real-world events caused the spread of these languages in order to create the aggregate relationships that we now observe? While such historical explanations are often *ad hoc*, this paper has attempted to explain synchronic variation using only empirical measures. While it is certainly the case that the concepts used here (predictive accuracy, region similarity, region uniqueness) tell us about varieties, it is not the case that they tell us the same things as traditional qualitative studies. In this case, two clear differences between this paper and traditional approaches to dialectology and dialectometry are (i) the focus on global variation with countries as the smallest spatial unit and (ii) the focus on written as opposed to spoken language.

First, we have a distinction between places (i.e., English used in the United States) and varieties (i.e., American English). There is a claim, whether implicit or explicit, in traditional dialectology that these two are not the same thing. For example, some speakers (older, male, rural, less educated) are taken as more representative than others (younger, urban, immigrant). A farmer born and raised in Kansas is assumed to be a local, a representative of American English; an IT specialist born in India but educated and living in Kansas is not. The argument in this paper, and perhaps in corpus-based research more broadly, is that this starting assumption is problematic. In short, we take American English to be English as used in the United States. We make no effort to exclude certain participants. This approach, then, can be situated within a larger movement away from NORM-based studies (Cheshire et al., 2015; Scherrer and Stoeckle, 2016).

Second, the dialect areas used in this paper ignore distinctions between native speakers and non-native speakers. Similar to the idea of locals vs. non-locals, the claim is that some places that produce a great deal of English data (for example, Nigeria or Malaysia) do not have the same status as American English as sources of ground-truth English data. This distinction is clearly a slippery-slope: while some language learners are not fully fluent, people who use a language like English for regular communicative functions cannot be categorized given *a priori* reasonings. We take this instead as an empirical question: language mapping is used to discover countries where English is regularly and robustly produced and dialect modeling is used to validate that these countries have distinct and predictable varieties. The social status of different English users (i.e., native vs. non-native) is entirely non-empirical and irrelevant. Given that these datasets do not come with individual demographics, however, it is important to also evaluate how well they reflect known demographic properties of the places they are taken to represent in order to ensure the connection between places and syntactic variants (Dunn and Adams, 2019).

Third, a distinction is sometimes made between varieties and dialects. For example, outer-circle and expanding-circle dialects are often called varieties. But what is the basis of this distinction? The argument in this paper is simple: the status of Nigerian English or Cameroon French or Angolan Portuguese is an empirical matter. The question is whether we can find these varieties using data-driven language mapping and can model their syntactic profile accurately enough to distinguish them from other varieties consistently across registers.

While previous work in dialectology and dialectometry focuses specifically on variation within individual countries, this paper has focused on global variation across many national varieties. One on the hand, this is important because the seven languages studied in this paper are used around the world: any local study will overlook important interactions. On the other hand, this means that these results are difficult to compare with previous small-scale studies. How could these methods be adapted to traditional problems of, for example, dividing Britain or the United States into dialect regions? First, there is no explicit spatial information provided to the models in this paper because the classes are all pre-defined. On approach would be to use existing sub-national administrative boundaries (such as postal codes) and apply a meta-classifier to evaluate different groupings. Which combinations lead to the highest predictive accuracy? This could be undertaken with the Twitter dataset but not with the web-crawled dataset.

4.2. Sociolinguistics and Computational Linguistics

Why should sociolinguistics more broadly care about a computational approach to dialectology? The first reason is simply a matter of descriptive adequacy: the models of variation in this paper have a broad and replicable feature space that is ultimately more meaningful and robust than multivariate models containing only a few features. While the grammars used are not explored further here, quantitative and qualitative evaluations are available elsewhere (Dunn, 2017, 2018a,b, 2019a). These models are more meaningful because they make predictions about categories as a whole (i.e., American English). They are more robust because they are evaluated against held-out samples using predictive accuracy. For both of these reasons, computational models of variation provide more accurate descriptions; this is important for quantitative sociolinguistics, then, simply as an extension of existing methods for discovering externally-conditioned variants (here, conditioned by geography). On the other hand, this approach of combining grammar induction and text classification produces models that, while easily understood in the aggregate, ultimately give us intricate and detailed descriptions that are difficult for human analysts to understand. The question is, do we expect human analysts to have full and complete meta-awareness for all variants in all national varieties of a language?

Beyond this, however, sociolinguistics is currently limited to small-scale studies, as discussed in the introduction. But the languages studied in this paper are used in many countries around the world. Each of these varieties has the potential to influence or be influenced by other distant varieties. In the same way, limiting a study to a handful of constructions ignores most of the functional capability of a language. Thus, current methods provide tiny snapshots of variation. But, moving forward, our ability to further understand syntactic variation and change depends on modeling entire grammars across all relevant varieties. While recent work has increased the number of features in order to produce larger-scale studies (Szmrecsanyi, 2013; Guy and Oushiro, 2015), such features remain language-specific and are defined *a priori*. On the other hand, however, a continued question for work that is bottom-up, such as this paper, is how to evaluate the connection between corpus-based models (which have been shown to be stable, robust, and highly accurate from an internal evaluation) and speech communities in the real world. How can computational descriptions and qualitative fieldwork be better combined?

Given the higher performance of lexical features in this paper, why should work in NLP that is not directly concerned with linguistic variation take a CxG or some other syntactic approach? There is an important distinction between topic variation (i.e., content arising from differences in human geography) and latent variation (i.e., structural variations arising from differences in variety). Any purely-lexical model is unable to distinguish between these two sources of information: Is this text written by someone from New Zealand or is it about New Zealand? Does this Tweet describe a vacation in New Zealand or was it written by a New Zealander on a vacation in the United States? Any model that is unable to distinguish between topical and latent properties within geo-referenced datasets will confuse these two types of cases. On the other hand, this is an incomplete approach the problem: how can we distinguish between topical variation, human geography-based varation, and linguistic variation within lexical items in order to have a better understanding of how these languages are used around the world? This remains a problem for future research.

Why should computational linguistics, and artificial intelligence more broadly, care about dialectology? As computational models become more important to society, it is essential that such models reflect all speakers equally. In spite of this, many models are biased against certain populations: either directly encoding the biases of individuals (Bolukbasi et al., 2016) or indirectly encoding a preference for dominant inner-circle varieties (Jurgens et al., 2017). Dialectometry can be used to prevent indirect biases against varieties like Nigerian English or Cameroon French by, first, identifying the relevant varieties that need to be considered and, second, providing a method to optimize language models for region-specific tasks. For example, if we can identify the membership of a sample that is part of an independent text classification problem (i.e., identifying helpful reviews or removing harrassing messages), then we can evaluate the degree to which existing models prefer dominant varieties (i.e., only suggesting reviews written in American English). This is important to ensure that inner-circle dominated training sets do not encode implicit biases against other varieties. It is also important because computational dialectometry can potentially improve equity between varieties in a way that traditional methods cannot.

DATA AVAILABILITY

The code and data for this paper can be found in the following locations:

- Construction Grammar package: https://github.com/jonathandunn/c2xg
- Common Crawl Data collection: https://github.com/jonathandunn/common_crawl_corpus
- Language identification code: https://github.com/jonathandunn/idNet
- Language identification models: https://labbcat.canterbury.ac.nz/download/?jonathandunn/idNet_models
- Common Crawl data: https://labbcat.canterbury.ac.nz/download/?jonathandunn/CGLU_v3
- Grammar Learning data: https://labbcat.canterbury.ac.nz/download/?jonathandunn/CxG_Data_FixedSize
- Experiment code, vectors, and raw results: https://labbcat.canterbury.ac.nz/download/?jonathandunn/Frontiers_in_AI

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The author would like to thank the editor, Jack Grieve, as well as the reviewers. This paper has also benefited from audiences at the University of Canterbury, NWAV-46, VarDial-19, LSA 2018 and from on-going conversations with Josh Trampier and Ben Adams.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2019.00015/full#supplementary-material

REFERENCES

Adams, B. (2015). Finding similar places using the observation-to-generalization place model. *J. Geograph. Syst.* 17, 137–156. doi: 10.1007/s10109-015-0209-3

Adams, B., and McKenzie, G. (2018). Crowdsourcing the character of a place: character-level convolutional networks for multilingual geographic text classification. *Trans. GIS* 22, 394–408. doi: 10.1111/tgis.12317

Argamon, S., and Koppel, M. (2013). A systemic functional approach to automated authorship analysis. *J. Law Policy* 12, 299–315.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web a collection of very large linguistically processed web-crawled corpora. *Lang. Resour. Eval.* 43, 209–226. doi: 10.1007/s10579-009-9081-4

Benko, V. (2014). "Aranea yet another family of (Comparable) web corpora," in *Proceedings of 17th International Conference Text, Speech and Dialogue* (Brno: Springer), 257–264.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). "Debiasing word embedding," in *30th Conference on Neural Information Processing Systems* (Barcelona), 1–9.

Calle-Martin, J., and Romero-Barranco, J. (2017). Third person present tense markers in some varieties of English. *Engl. World-Wide* 38, 77–103. doi: 10.1075/eww.38.1.05cal

Cheshire, J., Nortier, J., and Adger, D. (2015). "Emerging multiethnolects in Europe," in *Queen Mary's Occasional Papers Advancing Linguistics, Vol. 33* (London, UK), 1–27.

Chomsky, N. (1957). *Syntactic Structures*. Berlin: Mouton & Co.

Collins, P. (2012). Singular agreement in there existentials an intervarietal corpus-based study. *English World-Wide* 33, 53–68. doi: 10.1075/eww.33.1.03col

Cook, P., and Brinton, J. (2017). Building and evaluating web corpora representing national varieties of english. *Lang. Resour. Eval.* 51, 643–662. doi: 10.1007/s10579-016-9378-z

Davies, M., and Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *Engl. World-Wide* 36, 1–28. doi: 10.1075/eww.36.1.01dav

Donoso, G., and Sanchez, D. (2017). "Dialectometric analysis of language variation in Twitter," in *Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects, Vol. 4* (Valencia), 16–25. doi: 10.18653/v1/W17-1202

Dunn, J. (2017). Computational Learning of Construction Grammars. *Lang. Cogn.* 9, 254–292. doi: 10.1017/langcog.2016.7

Dunn, J. (2018a). Finding variants for construction-based dialectometry a corpus-based approach to regional CxGs. *Cogn. Linguist.* 29, 275–311. doi: 10.1515/cog-2017-0029

Dunn, J. (2018b). ‘Modeling the complexity and descriptive adequacy of construction grammars," in *Proceedings of the Society for Computation in Linguistics* (Salt Lake City, UT), 81–90.

Dunn, J. (2019a). "Frequency vs. association for constraint selection in usage-based construction grammar," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Minneapolis, MN: Association for Computational Linguistics).

Dunn, J. (2019b). "Modeling global syntactic variation in english using dialect classification," in *Proceedings of the NAACL 2019 Sixth Workshop on NLP for Similar Languages, Varieties and Dialects* (Minneapolis, MN: Association for Computational Linguistics), 42–53.

Dunn, J., and Adams, B. (2019). "Mapping languages and demographics with georeferenced corpora," in *Proceedings of Geocomputation 2019* (Queenstown), 16.

Dunn, J., Argamon, S., Rasooli, A., and Kumar, G. (2016). Profile-based authorship analysis. *Liter. Linguist. Comput.* 31, 689–710. doi: 10.1093/llc/fqv019

Eisenstein, J., O'Connor, B., Smith, N., and Xing, E. (2010). "A latent variable model for geographic lexical variation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Vol. 287 (Cambridge, MA: Association for Computational Linguistics), 221–227.

Eisenstein, J., O'Connor, B., Smith, N., and Xing, E. (2014). Diffusion of lexical change in social media. *PLoS ONE* 9:e113114. doi: 10.1371/journal.pone.0113114

Ginter, F., Hajič, J., and Luotolahti, J. (2017). *CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings*. Vancouver, BC: LINDATCLARIN digital library at the Institute of Formal and Applied Linguistics (FAL), Faculty of Mathematics and Physics, Charles University.

Goldberg, A. (2006). *Constructions at Work The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). "Building large monolingual dictionaries at the leipzig corpora collection from 100 to 200 languages," in *Proceedings of the Eighth Conference on Language Resources and Evaluation* (Istanbul: European Language Resources Association), 759–765.

Goldsmith, J. (2015). "Towards a new empiricism for linguistics," in *Empiricism and Language Learnability*, eds N. Chater, A. Clark, J. Goldsmith, and A. Perfors (Oxford: Oxford University Press), 58–105.

Grafmiller, J., and Szmrecsanyi, B. (2018). Mapping out particle placement in Englishes around the world A study in comparative sociolinguistic analysis. *Lang. Variat. Change* 30, 385–412. doi: 10.1017/S0954394518000170

Graham, S., Hale, S., and Gaffney, D. (2014). Where in the world are you? Geolocation and language identification on Twitter. *Profess. Geogr.* 66, 568–578. doi: 10.1080/00330124.2014.907699

Grieve, J. (2011). A regional analysis of contraction rate in written Standard American English. *Int. J. Corpus Linguist.* 16, 514–546. doi: 10.1075/ijcl.16.4.04gri

Grieve, J. (2012). A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguist. Linguist. Theory* 8, 39–72. doi: 10.1515/cllt-2012-0003

Grieve, J. (2013). A statistical comparison of regional phonetic and lexical variation in American English. *Liter. Linguist. Comput.* 28, 82–107. doi: 10.1093/llc/fqs051

Grieve, J. (2016). *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Lang. Variat. Change* 23, 1–29. doi: 10.1017/S095439451100007X

Grieve, J., Speelman, D., and Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *J. Linguist. Geogr.* 1, 31–51. doi: 10.1017/jlg.2013.3

Guy, G., and Oushiro, L. (2015). "The effect of salience on co-variation in Brazilian Portuguese," in *University of Pennsylvania Working Papers in Linguistics, Vol. 21* (Philadelphia, PA), 18.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Groningen: University of Groningen.

Hirst, G., and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Liter. Linguist. Comput.* 22, 405–417. doi: 10.1093/llc/fqm023

Hulden, M., Silfverberg, M., and Francom, J. (2015). "Kernel density estimation for text-based geolocation," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, TX: Association for the Advancement of Artificial Intelligence), 145–150.

Joachims, T. (1998). "Text categorization with support vector machines Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning* (Berlin: Springer), 137–142.

Jurgens, D., Tsvetkov, Y., and Jurafsy, D. (2017). "Incorporating dialectal variability for socially equitable language identification," in *Proceedings of the Annual Meeting for the Association for Computational Linguistics* (Vancouver, BC: Association for Computational Linguistics), 51–57.

Kachru, B. (1990). *The Alchemy of English The Spread, Functions, and Models of Non-native englishes*. Urbana-Champaign, IL: University of Illinois Press.

Kachru, B. E. (1982). *The Other Tongue: English across cultures*. University of Illinois Press, Urbana-Champaign.

Kondor, D., Csabai, I., Dobos, L., Szule, J., Barankai, N., Hanyecz, T., et al. (2013). "Using robust PCA to estimate regional characteristics of language-use from geotagged twitter messages," in *Proceedings of IEEE 4th International Conference on Cognitive Infocommunications* (Budapest: Institute of Electrical and Electronics Engineers), 393–398.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability Unmasking pseudonymous authors. *J. Mach. Learn. Res.* 8, 1261–1276.

Kretzschmar, W. A. (1992). Isoglosses and predictive modeling. *Amer. Speech* 67, 227–249. doi: 10.2307/455562

Kretzschmar, W. A. (1996). Quantitative areal analysis of dialect features. *Lang. Variat. Change* 8, 13–39. doi: 10.1017/S0954394500001058

Kretzschmar, W. A., Juuso, I., and Bailey, C. (2014). Computer simulation of dialect feature diffusion. *J. Linguist. Geogr.* 2, 41–57. doi: 10.1017/jlg.2014.2

Kroon, M., Medvedeva, M., and Plank, B. (2018). "When simple n-gram models outperform syntactic approaches discriminating between Dutch and Flemish," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects* (Santa Fe, NM), 225–244.

Kruger, H., and van Rooy, B. (2018). Register variation in written contact varieties of English A multidimensional analysis. *Engl. World-Wide* 39, 214–242. doi: 10.1075/eww.00011.kru

Labov, W., Ash, S., and Boberg, C. (2005). *The Atlas of North American English Phonetics, Phonology and Sound Change*. Berlin: De Gruyter Mouton.

Langacker, R. (2008). *Cognitive Grammar A Basic Introduction*. Oxford: Oxford University Press.

Lourentzou, I., Morales, A., and Zhai, C. (2017). "Textbased geolocation prediction of social media users with neural networks," in *Proceedings of 2017 IEEE International Conference on Big Data* (Boston, MA: Institute of Electrical and Electronics Engineers), 696–705.

Majliš, M., and Žabokrtský, Z. (2012). "Language richness of the web," in *Proceedings of the International Conference on Language Resources and Evaluation* (Istanbul: European Language Resources Association), 2927–2934.

Mocanu, D., Baronchelli, A., Perra, N., Goncalves, B., Zhang, Q., and Vespignani, A. (2013). The Twitter of Babel: mapping world languages through microblogging platforms. *PLoS ONE* 8:e61981. doi: 10.1371/journal.pone.0061981

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Liter. Linguist. Comput.* 21, 463–476. doi: 10.1093/llc/fql041

Nerbonne, J. (2009). Data-driven dialectology. *Lang. Linguist. Compass* 3, 175–198. doi: 10.1111/j.1749-818X.2008.00114.x

Nerbonne, J., and Kretzschmar, W. (2013). Dialectometry++. *Liter. Linguist. Comput.* 28, 2–12. doi: 10.1093/llc/fqs062

Rangel, F., Rosso, P., Potthast, M., and Stein, B. (2017). "Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in twitter," in *CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*, vol. 1866. Available online at: https://ceur-ws.org

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). "Supervised text-based Geolocation using Language Models on an Adaptive Grid," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, eds M. S. Stephen, S. Rallapalli, B. Wing, and J. Baldridge (Jeju-si: Association for Computational Linguistics), 1500–1510.

Ruette, T., and Speelman, D. (2014). Transparent aggregation of variables with individual differences scaling. *Liter. Linguist. Comput.* 29, 89–106. doi: 10.1093/llc/fqt011

Sanders, N. C. (2007). "Measuring syntactic difference in British English," in *Proceedings of the ACL 2007 Student Research Workshop*, Vol. 45, 1–6. doi: 10.3115/1557835.1557837

Sanders, N. C. (2010). *A statistical method for syntactic dialectometry*. (dissertation). Bloomington, IN, United States.

Scherrer, Y., and Stoeckle, P. (2016). A quantitative approach to Swiss German - Dialectometric analyses and comparison of linguistic levels. *Dial. Geolinguist.* 24, 92–125. doi: 10.1515/dialect-2016-0006

Schilk, M., and Schaub, S. (2016). Noun phrase complexity across varieties of English Focus on syntactic function and text type. *Engl. World-Wide* 37, 58–85. doi: 10.1075/eww.37.1.03sch

Skadiš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). "Billions of parallel words for free," in *Proceedings of the International Conference on Language Resources and Evaluation* (Reykjavik: European Language Resources Association).

Szmrecsanyi, B. (2009). Corpus-based dialectometry Aggregate morphosyntactic variability in British English dialects. *Int. J. Humanit. Arts Comput.* 2, 279–296. doi: 10.3366/E1753854809000433

Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, B. (2014). "Forests, trees, corpora, and dialect grammars," in *Aggregating Dialectology, Typology, and Register Analysis Linguistic Variation in Text and Speech* (Berlin: Mouton De Gruyter), 89–112.

Szmrecsanyi, B., Grafmiller, J., Heller, B., and Rothlisberger, M. (2016). Around the world in three alternations Modeling syntactic variation in varieties of English. *English World-Wide* 37, 109–137. doi: 10.1075/eww.37.2.01szm

Tamaredo, I. (2018). Pronoun omission in high-contact varieties of English Complexity versus efficiency. *English World-Wide* 39, 85–110. doi: 10.1075/eww.00004.tam

Tiedemann, J. (2012). "Parallel data, tools and interfaces in OPUS," in *Proceedings of the International Conference on Language Resources and Evaluation* (Istanbul: European Language Resources Association).

United Nations (2017). *World Population Prospects: The 2017 Revision, DVD Edition*. New York, NY: United Nations Population Division.

Wieling, M., and Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Comput. Speech Lang.* 25, 700–715. doi: 10.1016/j.csl.2010.05.004

Wieling, M., and Nerbonne, J. (2015). Advances in dialectometry. *Annu. Rev. Linguist.* 1, 243–264. doi: 10.1146/annurev-linguist-030514-124930

Wing, B., and Baldridge, J. (2014). "Hierarchical discriminative classification for text-based geolocation," in *Proceedings of the Conference on Empirical Methods in NLP* (Association for Computational Linguistics), 336–348.

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: measuring variation in the success of anglicisms in Dutch. *Cogn. Linguist.* 23, 749–792. doi: 10.1515/cog-2012-0023

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/frai.2019.00023

Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes

Benedikt Szmrecsanyi[1]*, Jason Grafmiller[2] and Laura Rosseel[3]

[1] Department of Linguistics, Faculty of Arts, KU Leuven, Leuven, Belgium, [2] Department of English Language and Linguistics, University of Birmingham, Birmingham, United Kingdom, [3] Linguistic and Literary Studies (LIST), Faculty of Arts and Philosophy, Vrije Universiteit Brussel, Brussels, Belgium

Inspired by work in comparative sociolinguistics and quantitative dialectometry, we sketch a corpus-based method (Variation-Based Distance & Similarity Modeling—VADIS for short) to rigorously quantify the similarity between varieties and dialects as a function of the correspondence of the ways in which language users choose between different ways of saying the same thing. To showcase the potential of the method, we present a case study that investigates three syntactic alternations in some nine international varieties of English. Key findings include that (a) probabilistic grammars are remarkably similar and stable across the varieties under study; (b) in many cases we see a cluster of "native" (a.k.a. Inner Circle) varieties, such as British English, whereas "non-native" (a.k.a. Outer Circle) varieties, such as Indian English, are a more heterogeneous group; and (c) coherence across alternations is less than perfect.

Keywords: comparative sociolinguistics, VADIS, probabilistic grammar, dialectometry, variationist linguistics

OPEN ACCESS

Edited by:
Tyler S. Kendall,
University of Oregon, United States

Reviewed by:
Kevin Tang,
Zhejiang University, China
Elisabeth Métais,
Conservatoire National des Arts et Métiers (CNAM), France
Jesse Egbert,
Northern Arizona University, United States

****Correspondence:***
Benedikt Szmrecsanyi
benedikt.szmrecsanyi@kuleuven.be

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *07 March 2019*
Accepted: *15 October 2019*
Published: *05 November 2019*

Citation:
Szmrecsanyi B, Grafmiller J and Rosseel L (2019) Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes. Front. Artif. Intell. 2:23. doi: 10.3389/frai.2019.00023

INTRODUCTION

Determining whether different varieties, dialects, or languages for that matter share the same or a similar "grammar" is an important and theoretically significant topic in comparative linguistics. In this paper we present a variationist method (Variation-Based Distance & Similarity Modeling—VADIS for short) to determine such similarity, based on naturalistic corpus and hence production data. VADIS builds bridges between subfields in sociolinguistics and variation studies that should be allied but that are in practice surprisingly disjoint. First, DIALECTOMETRY (see e.g., Séguy, 1971; Goebl, 1982; Nerbonne et al., 1999) is concerned with aggregate measures of linguistic similarity and distance as a function of geographic space; what is at issue is inter-speaker variation, where language users of dialect A use form X and language users of dialect B use form Y. Second, VARIATIONIST LINGUISTICS (see e.g., Labov, 1969; Gries, 2003; Bresnan et al., 2007) takes an interest in how speakers choose between formally distinct variants to express the same meaning, subject to probabilistic constraints that may be language-internal, stylistic, or language-external in nature; variationist linguistics, then, is in the first place all about intra-speaker variability (or "variability in the linguistic signal within a given language," in the parlance of van Hout and Muysken, 2016, p. 250), that is, variation between forms that are in principle available to all members of a given speech community. The basic idea behind VADIS is to use the output of variationist modeling as an input to dialectometric analysis, or—in other words—to measure inter-speaker variation by assessing the structure of intra-speaker variability.

Why do we need VADIS? There is, of course, an extensive literature on how to determine the grammatical similarity of varieties and dialects based on dialect atlases or survey data (for example, Spruit et al., 2009; Szmrecsanyi and Kortmann, 2009; Cysouw, 2013). Using naturalistic corpus data to measure the grammatical similarity of varieties is a trickier task. One avenue consists of establishing the text frequencies of forms and constructions in corpora, and to distill geolinguistic patterns from the frequency signal (Szmrecsanyi, 2013; Grieve, 2016). But VADIS digs even deeper than that: what counts is not if and/or how often people use particular constructions, but how they choose between "alternate ways of saying 'the same' thing" (Labov, 1972, p. 188). VADIS takes advantage of the fact that variationist analysis is good at quantifying the probabilistic grammar(s)—the set of constraints and their probabilistic effects on how people choose between variants of a particular variable[1]—of intra-speaker variation, and essentially defines the similarity between varieties as being proportional to how similar the probabilistic grammars regulating variation are. This is a more thoroughgoing, less "surfacy" method in comparison to the above-mentioned classical similarity-estimation methods: note that two dialects may have the exact same inventory of forms, and (though unlikely) these forms may even occur with the exact same text frequency—but still, the probabilistic conditioning of the forms may vary. VADIS is the only currently available method that will work under such circumstances.

VADIS builds on methods developed in comparative sociolinguistics (e.g., Tagliamonte, 2001), which has been used for decades to evaluate the relatedness of typically a small number of dialects drawing on multivariate evidence of typically a single variation phenomenon: are the same constraints significant across varieties? Do the constraints have similar effect sizes? Is the overall ranking of constraints similar? Unlike classical comparative sociolinguistics, however, VADIS scales up better to the study of a potentially infinite number of varieties based on many variation phenomena.

To showcase the descriptive and theoretical potential of the VADIS method, we analyze by way of a case study similarity patterns and relationships between varieties of English, fueled by a variationist analysis of three syntactic alternations:

(1) The genitive alternation (Heller et al., 2017)
  a. *the country's economic crisis* (the *s*-genitive)
  b. *the economic growth of the country* (the *of*-genitive)
(2) The dative alternation (Röthlisberger et al., 2017)
  a. *I'd given Heidi my T-Shirt* (the ditransitive dative variant)
  b. *I'd given the key to Helen* (the prepositional dative variant)
(3) The particle placement alternation (Grafmiller and Szmrecsanyi, 2018)
  a. *just cut the tops off* (verb-object-particle order)
  b. *cut off the flowers* (verb-particle-object order)

In principle, it is the analyst's decision which alternation(s) to include in the analysis; VADIS does not impose any restrictions, as long as linguistic choice-making can be modeled as a function of clearly defined language-internal and and/or language-external probabilistic constraints. In the case study at hand, the three alternations above were selected as they are all positional alternations subject to similar probabilistic constraints (e.g., constituent weight, constituent animacy, and so on).

The alternations in (1–3) are studied in nine World Englishes (British English, Canadian English, Irish English, New Zealand English, Hong Kong English, Indian English, Jamaican English, Philippine English, and Singapore English), based on materials from the International Corpus of English (ICE) and the Corpus of Global Web-Based English (GloWbE). Relevant observations of the (a) and (b) variants above were annotated for ∼10 probabilistic constraints including e.g., the principle of end weight (longer constituents tend follow shorter constituents; see e.g., Wasow and Arnold, 2003) and animacy effects (animate constituents tend to occur early; see e.g., Rosenbach, 2008).

Analysis indicates, among other things, that (a) probabilistic grammars are remarkably similar and stable across the varieties under study; (b) in many cases we see a cluster of "native" (a.k.a. Inner Circle) varieties, such as British English, whereas "non-native" (a.k.a. Outer Circle) varieties, such as Indian English, are a more heterogeneous group; and (c) coherence across alternations is less than perfect.

This paper is structured as follows: Section Data discusses the datasets we investigate. Section Spelling out the Variation-Based Distance & Similarity Modeling (VADIS) Method explains the VADIS method. In sections Quantification via similarity coefficients, Mapping out (dis)similarity relationships between varieties, and Assessing coherence, we present results. Section Discussion and Conclusion offers a discussion and conclusion.

DATA

In this paper, we re-analyze the genitive alternation dataset investigated by Heller (2018), the dative alternation dataset investigated by Röthlisberger (2018), and the particle placement dataset investigated by Grafmiller and Szmrecsanyi (2018) (see examples (1–3) above). The three datasets have been created in the context of the same project, and share the same basic design. With an interest in comparative probabilistic variation analysis, team members tapped into the International Corpus of English[2] (ICE) (Greenbaum, 1991) and the Corpus of Global Web-based English[3] (GloWbE) (Davies and Fuchs, 2015) to investigate syntactic variability in the following nine varieties of English:

- British English (henceforth: BrE)
- Canadian English (CanE)
- Irish English (IrE)
- New Zealand English (NZE)
- Jamaican English (JamE)

[1] The concept of a probabilistic grammar thus largely overlaps with what variationist sociolinguists refer to as a "variable grammar," defined by Tagliamonte (2006, p. 240), citing Poplack and Tagliamonte (2001, p. 91), as being represented by "the hierarchy of constraints constituting each factor [that regulates variation]".

[2] http://ice-corpora.net/ice/index.html

[3] https://www.english-corpora.org/glowbe/

- Singapore English (SgE)
- Indian English (IndE)
- Hong Kong English (HKE)
- Philippine English (PhlE)

ICE, initiated in 1990, is an ongoing project which was designed to create a set of parallel, balanced corpora representative of language usage across a wide range of (standard) national varieties. Each ICE component contains 500 texts of ∼2,000 words each, sampled from 12 spoken and written genres/registers. ICE components included here contain data from the early 1990s, with some also containing data collected as late as the early 2000s. Sampling for each national component is conducted by local teams following a common corpus design and annotation scheme to ensure maximal comparability across the components. GloWbE contains data collected from 1.8 million English language websites—both blogs and general web pages—from 20 different countries (∼1.8 billion words in all). To keep the datasets to a manageable size, texts were randomly sampled from each of the nine varieties in GloWbE, totaling 500,000 words per variety.

Areally, we are dealing with a convenience sample, subject to the limits of the availability of corpora. But a deliberate attempt was made to evenly balance what (e.g., Kachru, 1985, 1992) has called "Inner Circle" varieties of English (BrE, IrE, CanE, and NZE) and "Outer Circle" varieties of English (JamE, SgE, IndE, HKE, and PhlE). The distinction between Inner Circle and Outer Circle varieties is roughly equivalent to McArthur (1998) distinction between English as a Native Language (ENL) varieties (about communities "in which the language is spoken and handed down as the mother tongue of the majority of the population"; Schneider, 2011, p. 30), and English as a Second Language (ESL) varieties (about communities "in which English has been strongly rooted for historical reasons and assumes important internal functions (often alongside indigenous languages), e.g., in politics (sometimes as an official or co-official language), education, the media, business life, the legal system, etc."; Schneider, 2011, p. 30). We know from the literature (see Szmrecsanyi and Röthlisberger, 2019 for discussion) that this is a very important dialect-typological distinction in English linguistics.

The goal was to compile datasets amenable to variationist analysis. That means that in a first step interchangeable genitive, dative, and particle placement variants were defined which could be paraphrased by the competing variant with no semantic change. So, for example, (4a) can be paraphrased by (4b), which is why (4a) is a token that would have been included in the dataset, but (5a) cannot—in any of the varieties we study—be paraphrased by (5b), which is why (5a) is not a token that would have been included in the dataset

(4) a. *the speech of the president*
b. *the president's speech*
(5) a. three *liters of wine*
b.? *wine's three liters*

For reasons of space, we cannot review the definitions of the variable contexts in detail here; the reader is referred to the discussions in Heller (2018), Röthlisberger (2018), and Grafmiller and Szmrecsanyi (2018).

After all interchangeable variants were identified in the materials (dative alternation: $N = 13{,}171$; genitive alternation: $N = 13{,}798$; particle placement alternation: $N = 11{,}454$), each observation was annotated, manually or automatically, for a multitude of known and less-well known constraints on syntactic variation. For example, the principle of end-weight (Behaghel, 1909; Wasow and Arnold, 2003) predicts that in VO languages such as English, "heavy" constituents should follow "lighter" constituents. Thus, team members determined (a) the length of the possessor and possessum phrases in the genitive alternation (prediction: comparatively long possessors should favor the *of*-genitive, because the *of*-genitive places the possessor phrase after the possessum phrase), (b) the length of the recipient and theme phrases in the dative alternation (prediction: comparatively long recipients should favor the prepositional dative, because the prepositional dative places the recipient phrase after the theme phrase), and (c) the length of the direct object in the particle placement alternation (prediction: long direct objects favor verb-particle-object order, which places the direct object after the particle). Again, for reasons of space we cannot discuss the annotation procedure in detail; the reader is referred to Heller (2018), Röthlisberger (2018), and Grafmiller and Szmrecsanyi (2018).

SPELLING OUT THE VARIATION-BASED DISTANCE AND SIMILARITY MODELING (VADIS) METHOD

Overview

VADIS is designed to measure the (dis)similarity of grammars. Grammar is understood here as a set of probabilistic grammars (a.k.a. "variable grammars" in variationist sociolinguistics parlance) conditioning a set of $N \geq 1$ alternations or variation phenomena (a.k.a. "variables" in variationist sociolinguistics parlance). A probabilistic grammar specifies the set of constraints (a.k.a. predictors or "conditioning factors" in variationist sociolinguistics parlance) regulating a given alternation.

VADIS builds on methods developed in comparative sociolinguistics (see e.g., Tagliamonte, 2001, 2012, 162–173; Tagliamonte et al., 2016), which is a sub-discipline in variationist sociolinguistics that evaluates the relatedness between varieties and dialects based on how similar the conditioning of variation is in these varieties. Comparative sociolinguists rely on three what they call "lines of evidence" to determine relatedness:

1. Are the same constraints significant across varieties?
2. Do the constraints have the same strength across varieties?
3. Is the constraint hierarchy similar?

Similarity thus assessed is then often interpreted as historical and genetic relatedness. VADIS draws inspiration from this literature and adapts the comparative sociolinguistics method so that it can be applied to datasets sampling (a) more than a couple of dialects or varieties, and (b) more than one variation phenomenon at a time. This is accomplished through more rigorous quantification.

Let us illustrate by coming back to our case study, which covers three syntactic alternations in some nine regional varieties of English. Our point of departure is the view that the dative, genitive, and particle placement alternations are alternations between different forms that have the same meaning. We specifically consider each alternation as coming with its own probabilistic grammar, which regulates how people choose between variants. For example, Bresnan et al. (2007) is a seminal study that calculates regression models that predict how speakers of US American English choose between ditransitive (e.g., *I'd given Heidi my T-Shirt*) and prepositional dative variants (e.g., *I'd given my T-Shirt to Heidi*). According to the formula of model A (Bresnan et al., 2007; **Figure 4**), a non-given theme significantly decreases the odds that speakers will choose a prepositional dative variant by some 67% ($b = -1.1$), while an inanimate recipient significantly *in*creases the odds for a prepositional dative variant by a factor of about 12 ($b = 2.5$). These effects are part of the probabilistic grammar that regulates dative choice in spoken US American English, as sampled in the Switchboard corpus. But what would happen if we fitted a parallel model on data of, say, British English? Would we obtain a different model formula? Would the same constraints be significant? Would they have the same effect size? VADIS is a method to address these questions in a rigorously quantitative fashion. The basic idea behind VADIS is that similarity between varieties is proportional to how similar probabilistic grammars and model formulas are.

The VADIS Pipeline

Practically speaking, VADIS consists of the following steps:

Step 1: define, per alternation, the p most important constraints on variation. In the case study we are reporting here, we set $p = 8$[4] and so include the eight most important predictors (across all varieties) for each alternation[5]. We thus choose, in the case study at hand, to hold the number of constraints constant across alternations for the sake of maximum comparability, but we stress that in principle, the number of constraints do not need to be the same, considering that some alternations would naturally lend themselves to having more constraints than others, depending on the extent of previous research and the complexity of the factors at play. To identify the most important predictors, we fit conditional random forest models across all varieties (i.e., not accounting for variety differences) and created a global variable importance ranking of the predictors; we also consulted the extant literature on the alternations in question. Other ways to define predictor sets are certainly possible, but this task is best left to the VADIS user, not to the method itself. In the case of multi-level categorical predictors, we simplified to binary contrasts whenever possible. The predictor sets thus generated are reported in **Table 1**. We skip a detailed discussion of individual predictors and instead refer the reader to the publications where the annotation of predictors are discussed in detail.

Step 2: Fit a series of mixed-effects logistic regression models, one per variety and alternation. The response variable is variant choice (e.g., *s*-genitive vs. *of*-genitive), and the independent variables are the predictor sets identified in step 1. Note that, following Gelman (2008), all numeric variables in the model should be standardized and categorical variables should be centered. This approach allows direct comparison of the magnitudes of the coefficients in the model. We use mixed-effects models (R function glmer()) with random intercepts for speaker/writer (approximated by corpus file id) and genre. Additional random intercepts were possessor and possessum head for the genitive alternation, verb and theme head for the dative alternation and particle verb and head of the direct object for the particle placement alternation. In previous studies, from which these data were taken, random slopes for a number of predictors were initially tried and evaluated. In most cases, models failed to converge, and in those that were successful, the random slopes were not statistically justified. In our experience, this is quite common with corpus-based grammatical alternation studies, where the individual group levels of the random effects (typically texts and/or lexical items) tend to be sparsely populated. There is also growing evidence that imposing maximal random effects structure where it is not supported can adversely affect results (Bates et al., 2015; Matuschek et al., 2017). Therefore we did not include random slopes for this study. The resulting models are of satisfactory quality: concordance statistic (C) values[6] are consistently greater than 0.88, and VIFs never exceed 2.5.

Step 3: Based on the variety-specific regression models, determine cross-variety similarity based on predictor significance[7]. In this step, we define the probabilistic distance between two varieties as being proportional to the extent to which the varieties do *not* overlap with regard to which constraints significantly (in the case study at hand, we set alpha $= 0.05$[8]) regulate variant choice. To exemplify, consider two

[4] We experimented with predictor sets of different sizes, from $p = 5$ to $p = 10$. In principle, larger predictor sets are preferable to smaller predictor sets, but then again including too many predictors that turn out as insignificant in many cases is problematic. Given these principles $p = 8$ seemed like a good compromise for the case study we report here. See Tamaredo et al. (2019) for a VADIS analysis that uses $p = 5$.

[5] The method as outlined here does not distinguish between different types of constraints, e.g., between what Tamminga et al. (2016, p. 303) term sociostylistic factors (*s*-conditioning), internal linguistic factors (*i*-conditioning), and physiological and psycholinguistic factors (*p*-conditioning). Note however that the method can be easily adapted to restrict attention to only particular types of constraints.

[6] The concordance statistic (or index) represents the probability that the model will rank any randomly chosen observation of the predicted variant higher than any randomly chosen observation of the alternate variant. C is equal to the area under the receiver operating characteristic curve. Note that model fit only matters for VADIS to the extent that the model fits are acceptable and reasonably close to one another across the same alternation. One probably should not compare models with C values of 0.75 and 0.95, but a range of 0.02 or 0.03 seems perfectly reasonable.

[7] We acknowledge that this step relies on null hypothesis significance testing based on ultimately arbitrary alpha levels, which is increasingly controversial. Note, however, that VADIS also includes two other lines of evidence which are more nuanced. The main reason why we include step 3 is that checking significances is a customary line of evidence in classical comparative sociolinguistics, and so for the sake of continuity with the extant literature VADIS also considers this criterion.

[8] A Bonferroni correction could in principle be used to make the alpha level more conservative, but we refrain from doing so here since our main interest lies with comparative analysis (using significance as an auxiliary criterion), and not with statistical significance *per se*.

TABLE 1 | Predictor sets used for the analysis.

Genitive alternation (see Heller et al., 2017)	**Dative alternation (see Röthlisberger et al., 2017)**	**Particle placement alternation (see Grafmiller and Szmrecsanyi, 2018)**
Possessor animacy (animate vs. inanimate)	Log weight ratio between recipient and theme	Length of the direct object in words
Possessor length in words	Recipient pronominality (pronominal vs. non-pronominal)	Definiteness of the direct object (definite vs. indefinite)
Possessum length in words	Theme complexity (complex vs. simple)	Givenness of the direct object (given vs. new)
Possessor NP expression type (NP vs. NC vs. other)	Theme head frequency	Concreteness of the direct object (concrete vs. non-concrete)
Final sibilancy in possessor (present vs. absent)	Theme pronominality (pronominal vs. non-pronominal)	Thematicity of the direct object
Previous choice (*of* vs. *s* vs. none)	Theme definiteness (definite vs. indefinite)	Directional modifier (present vs. absent)
Semantic relation (prototypical vs. non-prototypical)	Recipient givenness (given vs. new)	Semantics (compositional vs. non-compositional)
Possessor head frequency	Recipient head frequency	Surprisal.P

hypothetical varieties A and B and five constraints a-e which regulate some variation phenomenon:

	Variety A	**Variety B**
Constraint a	Significant	Significant
Constraint b	Significant	Not significant
Constraint c	Not significant	Significant
Constraint d	Not significant	Not significant
Constraint e	Significant	Significant

Variety A and B agree on the significance of three constraints (a, d, e), and disagree with regard to two constraints. The distance between the two varieties is thus two out of five squared Euclidean distance points. Scaling this to an interval between 0 (no disagreement whatsoever) and 1 (maximal disagreement) yields, in the fictitious example at hand, a distance value of 2/5 = 0.4 and a corresponding similarity value of 3/5 = 0.6.

Step 4: Based on the variety-specific regression models, determine cross-variety distance and similarity based on the magnitude of effects. To define the similarity between the varieties, this step compares the extent to which the effect sizes of the constraints in the various regression models are similar (inspired by the procedure sketched in Heller, 2018). This is done by calculating a distance matrix based on the model estimates (using Euclidean distance), whether or not they are significant[9].This is illustrated with a toy example in **Tables 2**, **3**. **Table 2** shows the model estimates of five constraints for three varieties. The Euclidean distances between these varieties, based on the estimates from **Table 2**, are presented in **Table 3**. The next step for this line of evidence is to calculate the mean distance per variety, i.e., the average of the pairwise distances between the varieties (cf. **Table 4**). To scale the distances to an interval between 0 and 1, we can ask the following question: what is the maximal distance between the varieties under study? We define this maximal distance here as the distance between two hypothetical varieties whose constraints have exactly the opposite effects. Such cases of complete constraint "flipping", i.e., a systematic reversal in the direction of *every constraint's effect* between two varieties, are very unlikely to happen in real world contexts. We set the absolute size of all the constraints to a reasonable value (±1) to create two (hypothetical) varieties that are about as different from one another as we could realistically expect two related varieties to be. For the toy case involving 5 constraints in **Table 2**, the maximum distance is calculated to be 4.47. We divide the observed distances by this value to give normalized distances within a range of 0 to 1. For the similarity scores we subtract these scaled distances from 1 to give us a score where larger values represent greater average similarity (cf. **Table 4**). Averaging over the similarities in our toy example gives a similarity coefficient of 0.42.

TABLE 2 | Model estimates for three fictitious varieties A, B, and C.

	Variety A	**Variety B**	**Variety C**
Constraint	−2.10	−1.50	1.20
Constraint	−1.30	−1.60	−1.20
Constraint	0.75	−0.05	0.63
Constraint	0.69	0.80	2.20
Constraint	−0.92	−1.0	−0.79

TABLE 3 | Distance matrix for fictitious varieties A, B, and C (Euclidean distance).

	Variety A	**Variety B**	**Variety C**
Variety A	0		
Variety B	1.05	0	
Variety C	3.63	3.15	0

Step 5: Fit a series of conditional random forest models, one per variety and alternation. To independently estimate the relative importance of the constraints, we use permutation-based variable importance rankings derived from conditional random forests (CRFs; Strobl et al., 2009). Like regression models, random forests are a supervised learning method

[9] A disadvantage of including all estimates in the model, also the ones of constraints that do not reach significance, is that the latter may not be very reliable. However, we have opted not to use significance as an arbitrary cut-off point in this line of evidence in order not to repeat the weakness of the first line (see also footnote 7 in that respect).

TABLE 4 | Mean distances and mean similarities per variety.

Variety	Mean distance	Mean distance (scaled)	Mean similarity
Variety B	2.10	0.47	0.53
Variety A	2.34	0.52	0.48
Variety C	3.39	0.76	0.24
Mean	2.61	0.58	0.42

that aims to predict an outcome from a set of predictor values, however, this is where the similarities end. Random forests are a decision tree-based ensemble method which offers various advantages over regression models. Random forests are more reliable with unbalanced data, and offer methods for assessing the conditional importance of individual predictors in CRFs. Additionally, cross-validation is built into the method, resulting in greater accuracy and more reliable importance measures. For these reasons we believe CRFs offer a valuable independent assessment of the relationship between the alternations and their constraints. For calculating the CRFs and variable importances we use the cforest() and varimpAUC() functions in R's party package[10]. The response variable and independent variables in the models are the same as for the regression models in step 2 (though inputs are not standardized for the CRFs)[11].

Step 6: Based on the variety-specific conditional random forest models, determine cross-variety distance and similarity based on the importance rankings of the predictors. In this last step, we measure the probabilistic distance between two varieties simply as the Spearman rank correlation between those varieties' respective variable importance rankings[12]. For example, consider the three hypothetical varieties A, B, and C with the constraint rankings below:

	Variety A	Variety B	Variety C
Constraint a	1	1	2
Constraint b	2	3	4
Constraint c	3	2	3
Constraint d	4	4	1
Constraint e	5	5	5

Varieties A and B show the greatest degree of similarity, with a correlation of $\rho = 0.9$, while varieties A and C are least similar, with a correlation of $\rho = 0.3$. Variety B is slightly more similar to variety C than variety A is ($\rho = 0.4$), but it is far more similar to A than to C. We can arrange these pairwise correlations in a table like so:

	Variety A	Variety B	Variety C
Variety A	1	0.9	0.3
Variety B	0.9	1	0.4
variety C	0.3	0.4	1

From the workflow described above, it is clear that the case study reported in this paper (analyzing the similarity of nine varieties based on three alternations, including various subsets of the data) generated hundreds of regression and CRF models. Hence, it is not possible to report a comprehensive overview of model quality measures for the case studies. Instead, we restrict ourselves reporting the C values for the regression models based on all available data in **Table 5** below.

An R package (under development) which performs all the above calculation is available at https://github.com/jasongraf1/VADIS. The analysis scripts we used to conduct our case study are available at https://osf.io/3gfqn/, along with the genitive alternation and dative alternation datasets (the particle placement dataset is built into the R package mentioned above).

About Concept Validity and Reliability

Given the novelty and complexity of the VADIS methodology, some evaluation of the method's validity and reliability is warranted. Preliminary work suggests that the similarity coefficients do indeed accurately and consistently capture relative degrees of similarity among varieties. In a study using a series of simulated datasets, designed with varying degrees of similarity, Heller (2018, p. 199–204) showed that the similarity coefficients derived from models fit to these datasets correlated inversely with the degree of variability built into the data simulation. The more variable the datasets were designed to be when they were created, the lower the similarity coefficients were for all three lines of evidence. In a second study, Röthlisberger (2018, p. 175; 215–216) used a bootstrapping procedure to assess the reliability of the similarity coefficients for each line of evidence across 1,000 bootstrap samples of her datives dataset. She found a high degree of consistency for all three lines of evidence with the second line (coefficient strength) being the most consistent and the third line (constraint ranking) being the least consistent. Finally, we assessed the validity of methods for visualizing similarities (visualization and mapping is discussed in section Mapping Out (dis)Similarity Relationships Between Varieties) via a second simulation study in which artificial datasets were constructed to vary in specific ways and then subjected to VADIS analysis. Results of the visualizations were exactly as predicted, e.g., datasets that were designed to have opposite constraint effects were maximally distinguished, while datasets designed to have nearly identical constraint effects clustered tightly together. In all, we conclude that the procedure is quite robust.

[10] The number of trees in the forests was set to 500, and the number of predictors sampled ("mtry") was set to 3. All other hyperparameters were left at the default values for the package functions.

[11] Note that no random effects were included given that mixed effects random forests are not yet fully implemented for classification problems.

[12] We stress that this measurement is only about the ranking of the constraints, and does not take graded differences in terms of the actual variable importance scores into account (see Strobl et al., 2009, p. 336 on why variable importance scores should not be directly compared across models). Graded differences are anyway covered by the 2nd line of evidence (step 4).

TABLE 5 | *C* values for glmer models and CRFs based on all available data.

	Dative alternation		Genitive alternation		Particle placement alternation	
	Glmer model	CRF	Glmer model	CRF	Glmer model	CRF
BrE	0.95	0.95	0.91	0.93	0.89	0.91
CanE	0.96	0.95	0.92	0.93	0.91	0.91
HKE	0.95	0.94	0.92	0.92	0.90	0.93
IndE	0.96	0.96	0.92	0.93	0.88	0.93
IrE	0.95	0.95	0.90	0.92	0.89	0.91
JamE	0.97	0.96	0.92	0.93	0.88	0.93
NZE	0.95	0.94	0.91	0.92	0.91	0.92
PhlE	0.96	0.97	0.90	0.91	0.89	0.94
SgE	0.95	0.95	0.91	0.92	0.91	0.93

QUANTIFICATION VIA SIMILARITY COEFFICIENTS

One way in which VADIS can address the issue of variation-based similarities consists of calculating what we will call here SIMILARITY COEFFICIENTS. The idea is to quantify the similarity between varieties by coefficients which range between 0 and 1, where 0 indicates total dissimilarity and 1 indicates total similarity. Similarity coefficients are calculated as follows: for every variation phenomenon under study, we obtain $n \times (n{-}1)/2$ unique pairwise similarity values for each line of analysis (steps 3, 4, and 6), where n is the number of varieties under analysis. For example, if we study, say, the dative alternation in 9 varieties, then we obtain $9 \times 8/2 = 36$ pairwise similarity values for each of the three lines of evidence. Subsequently, we calculate one mean similarity coefficient per line of evidence by simply taking the arithmetic mean of all pairwise similarity values. In the case study at hand with 9 varieties of English, this means that each of the similarity coefficients averages over 36 pairwise similarity values.

Table 6 displays similarity coefficients across lines of evidence and alternations, based on all available data and including all nine regional varieties of English under study. The coefficients range between 0.46 (2nd line, particle placement alternation) and 0.83 (3rd line, genitive alternation). The last row displays mean similarity coefficients per alternation across lines of evidence. So the mean similarity coefficient for the genitive alternation is 0.74; for the dative alternation it is 0.64; and for the particle placement alternation it is 0.68. In other words, the genitive alternation is most stable across varieties, and the dative alternation is least stable; the particle placement alternation takes the middle road. As far as the three different lines of evidence are concerned, we note that the 1st line (significance) and the 3rd line (constraint ranking) yield on average similarly sized coefficients; 2nd line measurements (effect strength) are substantially lower in the case of the genitive and dative alternations, though not in the particle placement alternation.

The value in the bottom row of the rightmost column of **Table 6** is what we would like to call the CORE GRAMMAR SCORE (Γ): it is the mean similarity coefficient across all alternations subject to study and thus abstracts away from particular alternations. In the case study at hand (3 syntactic alternations × 9 varieties of English; all available data), we obtain a core grammar score of $\Gamma = 0.69$. Relying on customary schemes for interpreting (correlation) coefficients (e.g., De Vaus, 2002, p. 272), we thus see "substantial to very strong" similarities between the varieties under study.

TABLE 6 | Similarity coefficients across lines of evidence and alternations.

	Genitive alternation	Dative alternation	Particle alternation	
1st line (significance)	0.81	0.68	0.73	
2nd line (effect strength)	0.60	0.46	0.69	
3rd line (ranking)	0.83	0.78	0.62	
mean	0.74	0.64	0.68	$\Gamma = 0.69$

Input dataset: all available data. Coefficients range between 0 (total dissimilarity) and 1 (total similarity).

The foregoing analysis is based on all available data. What would happen if we restricted attention to particular subsets of the data? **Table 7** reports core grammar scores Γ for a number of sub-datasets, along with hierarchies of stability as far as individual alternations are concerned. When VADIS is run on particular sub-datasets (as opposed to the full dataset), then, core grammar scores tend to be higher, thanks to the fact the sub-datasets in question are by definition more homogeneous (spoken only, Inner Circle only, etc.) The largest core grammar score is obtained when attention is restricted to Inner Circle varieties ($\Gamma = 0.80$), indicating that these varieties are particularly homogeneous and similar to each other. Outer Circle varieties are substantially less homogeneous, with a core grammar score of $\Gamma = 0.73$. As to the difference that medium makes, written varieties are somewhat more homogeneous ($\Gamma = 0.75$) than spoken varieties ($\Gamma = 0.72$). Turning to differences between alternations, we have seen before that when we investigate all available data, the hierarchy of stability is genitives > particles > datives (meaning that the way language users choose between genitive variants is most similar across varieties, while dative choices are least similar). The genitive alternation turns out to be most stable also when we restrict attention to various sub-datasets, with the exception of the spoken sub-dataset, where the genitive alternation is actually the

TABLE 7 | Core grammar scores (Γ) and hierarchies of stability for subsets of the data.

	Core grammar score (Γ)	Hierarchy of stability
All available data (**Table 6**)	$\Gamma = 0.69$	Genitives > particles > datives
Spoken data only (ICE-s)	$\Gamma = 0.72$	Datives > particles > genitives
Written data only (ICE-w and GloWbE)	$\Gamma = 0.75$	Genitives > datives > particles
Inner Circle varieties only (BrE, IrE, CanE, NZE)	$\Gamma = 0.80$	Genitives > particles > datives
Outer Circle varieties only (HKE, SgE, IndE, JamE, PhlE)	$\Gamma = 0.73$	Genitives > datives > particles

least stable one. This is primarily due to a very low similarity coefficient (0.37) for the 3rd line of evidence in spoken materials, meaning that the rankings of constraints on genitive variation are rather dissimilar across varieties.

MAPPING OUT (DIS)SIMILARITY RELATIONSHIPS BETWEEN VARIETIES

We have seen in the preceding section how VADIS yields similarity coefficients to precisely quantify the (dis)similarity between regionally specific probabilistic grammars. In the case study we have investigated, we have seen that the similarity coefficients tend toward the similarity pole—for example, the core grammar score calculated on the basis of all available data came out at $\Gamma = 0.69$ (again, on a scale between 0—indicating maximal dissimilarity—and 1—indicating maximal similarity). So there is clearly more similarity than dissimilarity, but crucially core grammar scores are mean values, and (dis)similarities are not necessarily evenly spread across the network of varieties under study. In this section we will demonstrate how VADIS can be used to visually depict (dis)similarity relationships between varieties.

The aim, then, is not to calculate *mean* similarity coefficients, but to arrange *pairwise* similarity coefficients in so-called distance matrices. Distance matrices are the customary input in classical dialectometry (Séguy, 1971; Goebl, 1982; Nerbonne et al., 1999; Szmrecsanyi, 2013) and work essentially like distance tables in road atlases, which specify geographic distances between locations. Let us illustrate drawing on our case study: for each alternation and each of the three lines of evidence, we create one distance matrix. We are exploring $n = 9$ varieties of English, which yields $n \times (n{-}1)/2 = 9 \times 8/2 = 36$ unique variety pairings. To each pairing, we assign the relevant inverse similarity coefficient (1—similarity coefficient), thus converting similarity coefficients into *dis*similarity values[13].

	BrE	CanE	HKE	IndE	IrE	JamE	NZE	PhlE
CanE	0.000							
HKE	0.310	0.310						
IndE	0.548	0.548	0.238					
IrE	0.286	0.286	0.048	0.167				
JamE	0.095	0.095	0.262	0.452	0.262			
NZE	0.095	0.095	0.190	0.476	0.167	0.048		
PhlE	0.286	0.286	0.452	0.571	0.333	0.405	0.310	
SgE	0.214	0.214	0.310	0.429	0.167	0.286	0.167	0.095

FIGURE 1 | VADIS distance matrix for the 3rd line of evidence in the particle placement alternation (all data included). Scores range between 0 (maximal similarity) and 1 (maximal dissimilarity).

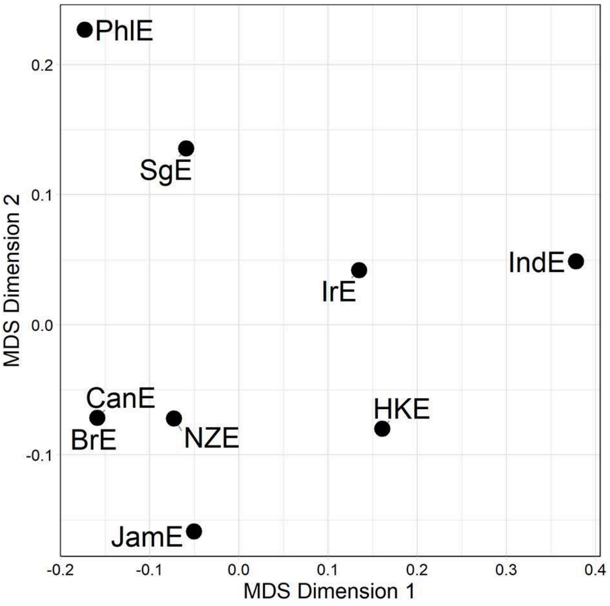


FIGURE 2 | MDS representation of 3rd line distances for the particle placement alternation (see **Figure 1**). Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

Figure 1 exemplifies by displaying the distance matrix for the 3rd line of evidence (constraint ranking) in the particle placement alternation. All distances are scaled between 0 (no distance) and 1 (maximal distance). Consider now e.g., the pairing between BrE and NZE, which is associated with a comparatively small distance value of 0.095. This is another way of saying that the similarity coefficient associated with this pairing is 1–0.095 = 0.905. In plain English, BrE, and NZE are very similar in terms of the constraint ranking in the particle placement alternation. By contrast, the distance between BrE and IndE is 0.548, which is considerably larger.

Distance matrices are informative but somewhat hard to process via eye balling. But there are a number of techniques in the dialectometric toolbox to visualize distance matrices. One of these is Multidimensional Scaling (MDS) (see e.g., Kruskal and Wish, 1978), which reduces a higher-dimensional distance matrix

[13] The distances that are calculated in VADIS are transitive, thus if the distance between variety A and B is 0, and the distance between B and C is 0, then the distance between variety A and C will also be 0.

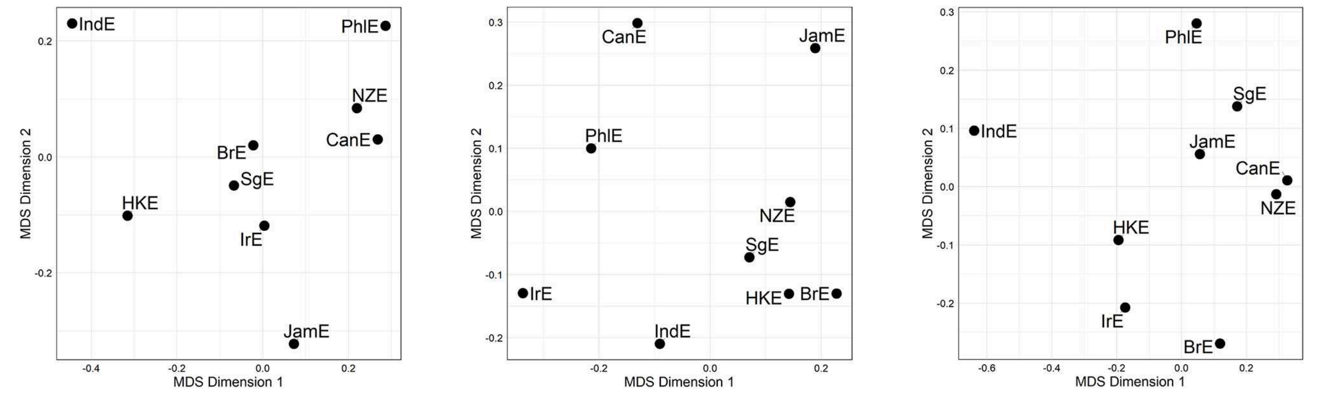


FIGURE 3 | MDS representation of compromise distances per alternation. **(Left)** genitive alternation. **(Middle)** dative alternation. **(Right)** particle placement alternation. Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

to a lower-dimensional representation which is more amenable to visualization[14]. The task before us here is to scale down the $n-1$ dimensional distance matrix (in which each of the nine varieties under study is characterized by its distance to the other eight varieties in the matrix) to a two-dimensional representation. Per alternation, we are initially dealing with three separate distance matrices (one per line of evidence), which could in principle be plotted separately. For example, **Figure 2** is a MDS representation of the distance matrix shown in **Figure 1**. Proximity in the plot is proportional to linguistic similarity. BrE and NZE are close in the plot, while BrE and IndE are fairly distant—which is of course in line with the numerical values in **Figure 1**.

Let us now abstract away from individual lines of evidence by fusing the three line-specific distance matrices, thus arriving at line-merged but alternation-specific distance matrices[15]. **Figure 3** displays the corresponding MDS plots. Cursory inspection of the plots reveals substantial differences between alternations (we will come back to this issue in the next section), but also similarities—for instance, across all three alternations, IndE and PhlE are at the periphery.

We may now take a further aggregation step for the sake of raising the analysis of (dis)similarity relationships to an even higher level of generalization. This we can accomplish by fusing the three alternation-specific-distance matrices (visualized in **Figure 3**) into a single compromise distance matrix merged across all lines and alternations, or Γ-MATRIX for short. An MDS visualization of this Γ-matrix is shown in **Figure 4**. In the plot, all Inner Circle varieties are clustered in the top right-hand quadrant, with SgE—which according to the literature is an Outer Circle variety in the process of becoming an Inner Circle variety (Leimgruber, 2013, p. 122)—forming part of that cluster. IndE and PhlE are outliers. Supplementary inspection of silhouette widths in hierarchical agglomerative cluster analysis (Levshina, 2015, p. 312) indicates that the distance matrix underlying **Figure 4** lacks substantial cluster structure.

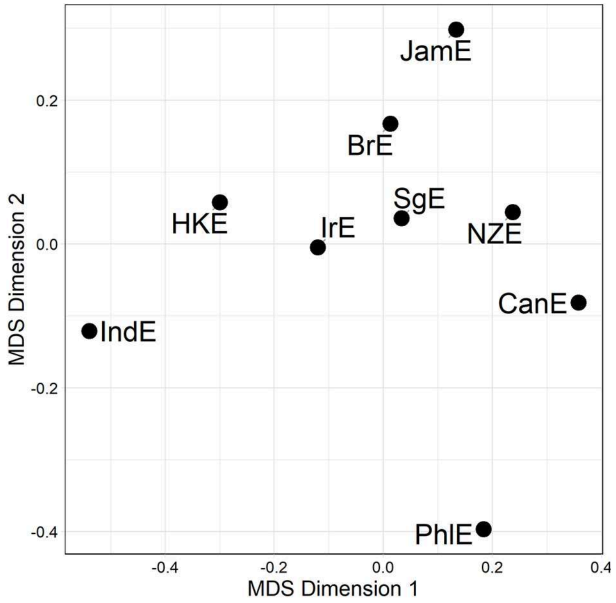


FIGURE 4 | MDS representation of the Γ-matrix (a single compromise distance matrix merged across all lines and alternations). Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

ASSESSING COHERENCE

Using the VADIS method means taking a lot of measurements. This section will discuss the extent to which these various measurements overlap with each other. We begin by exploring coherence between the three lines of evidence (constraint significance, constraint strength, and constraint ranking). The question is if large differences between any two varieties

[14] In this study, we are using R's cmdscale() function to obtain MDS solutions.

[15] We use the fuse() function in R package analogue to fuse distance matrices (see https://cran.r-project.org/web/packages/analogue/analogue.pdf). All input matrices are weighted equally. This could in principle be changed, but we see no compelling reason to weigh up or down particular lines of evidence.

TABLE 8 | Mantel correlation coefficients between distance matrices, based on all available data.

	Genitive alternation	Dative alternation	Particle alternation
Overlap 1st line/2nd line	***r* = 0.41** **(*p* = 0.03)**	*r* = 0.12 (*p* = 0.34)	***r* = 0.36** **(*p* = 0.05)**
Overlap 1st line/3rd line	*r* = 0.07 (*p* = 0.36)	*r* = −0.01 (*p* = 0.50)	*r* = 0.25 (*p* = 0.13)
Overlap 2nd line/3rd line	***r* = 0.47** **(*p* = 0.03)**	*r* = −0.15 (*p* = 0.77)	***r* = 0.68** **(*p* = 0.00)**

Significant coefficients are bolded.

TABLE 9 | Mantel correlation coefficients between fused distance matrices (combining all lines of evidence and based on all available data).

Overlap genitive alternation/dative alternation	*r* = 0.05 (*p* = 0.41)
Overlap genitive alternation/particle alternation	***r* = 0.52 (*p* = 0.01)**
Overlap dative alternation/particle alternation	*r* = 0.11 (*p* = 0.31)

Significant coefficients are bolded.

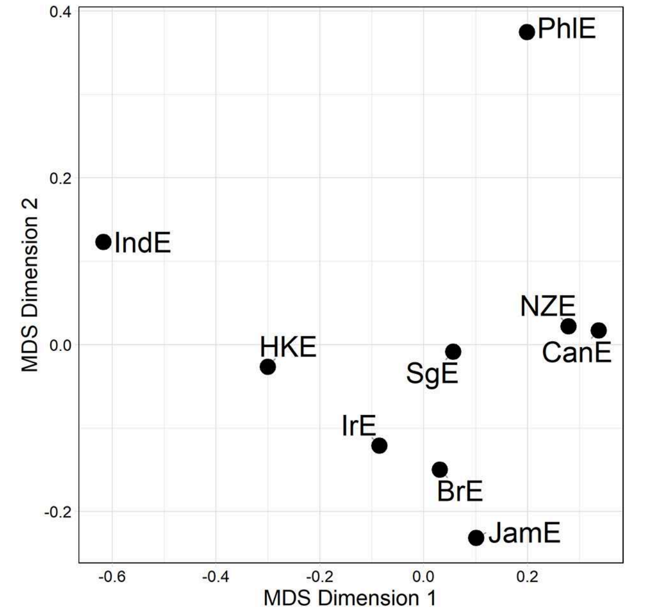


FIGURE 5 | MDS representation of a compromise distance matrix merged across the genitive and particle placement alternation (all available data). Distances between data points in plot is proportional to probabilistic grammar distances between varieties.

according to one particular line of evidence predict large differences between the same two varieties also according to the other lines of evidence. To exemplify, let us re-consider the distance matrix in **Figure 1**, which is about distances between varieties according to the 3rd line of evidence (constraint ranking) in the particle placement alternation. **Figure 1** showed that according to the 3rd line of evidence, BrE and NZE are comparatively close linguistically, while BrE and IndE are comparatively distant. The question is if BrE and NZE will also turn out as close, and BrE and IndE as distant, according to the other lines of evidence.

We measure overlap between distance matrices using the Mantel test (Levshina, 2015, p. 348–349), which, based on permutation, yields correlation coefficients that range between 0 (no overlap) and 1 (total overlap)[16]. **Table 8** displays the results. Observe, first, that the dative alternation is the odd one out in that none of the lines overlap with each other in this alternation. Second, the genitive alternation and the particle placement alternation are similar in that they both show moderate but significant overlap between the first line of evidence (constraint significance) and the second line of evidence (constraint strength), as well as substantial overlap between the second line of evidence and the third line of evidence (constraint ranking). We do not see significant overlap anywhere between the first line of evidence and the third line of evidence.

A related issue concerns the overlap, or coherence, between different alternations. We are concretely asking the following question: if, according to alternation A, two varieties are close in terms of how people choose between different ways of saying the same thing, will the two varieties also turn out to be close when the analysis is based on alternations B and C? Again, we turn to calculating Mantel coefficients between the relevant distance matrices (**Table 9**).

The upshot is, then, that there is significant and substantial overlap between the genitive alternation and the particle placement alternation, while the dative alternation does not overlap with either one of the other alternations. Against this backdrop, it is instructive to combine the genitive and particle placement alternation-based distance matrices—given their overlap—without throwing the dative distance matrix into the mix. **Figure 5** shows an MDS representation of this combined genitive/particle placement distance matrix.

The pattern in **Figure 5** is that the Inner Circle varieties are clustered in the lower right-hand quadrant in **Figure 5**; this quadrant also contains JamE and SgE. PhlE and IndE are outliers. Compare this to the dative alternation-only plot (middle plot **Figure 3**), from which no discernible pattern arises at all.

DISCUSSION AND CONCLUSION

Drawing inspiration from comparative sociolinguistics and dialectometry, we have sketched in this paper a method—Variation-Based Distance & Similarity Modeling (or VADIS for short)—that gauges the extent and structure of inter-speaker variation through assessing intra-speaker variation. VADIS specifically estimates the similarity between varieties and dialects as a function of how similar the ways are in which language users choose between different ways of saying the same thing. On the technical plane, VADIS calculates a series of multivariate models that predict speakers' and writers' linguistic choices, and utilizes three criteria to calculate similarity and distance measures: (1)

[16] We use the mantel() function in R package vegan to calculate Mantel coefficients (see https://cran.r-project.org/web/packages/vegan/vegan.pdf).

Are the same constraints significant across varieties? (2) What is the extent to which constraints have similar effect strengths? (3) What is the extent to which the ranking of constraints is similar? With its focus on how people make choices and thanks to its reliance on naturalistic corpus data as data source, VADIS has a more usage-based bent than classical dialectometry, and is able to pick up differences even in cases where varieties happen to have the same inventory of forms and exhibit similar frequencies, but with possibly different underlying probabilistic grammars. We noted also that the quantitative rigor of VADIS scales up better to more varieties and more variation phenomena than classical comparative sociolinguistics.

To illustrate how VADIS can characterize (dis)similarities across and relationships between varieties, we presented a case study about three syntactic alternations (the genitive alternation, the dative alternation, and the particle placement alternation) in nine World Englishes, four of which are Inner Circle, or English-as-a-native-language, varieties (BrE, CanE, IrE, and NZE), and five of which are Outer Circle, or English-as-a-second-language, varieties (IndE, HKE, SgE, PhlE, and JamE). Key findings uncovered through VADIS may be summarized as follows.

First, we showed in section Quantification via Similarity Coefficients how VADIS can precisely quantify, via similarity coefficients, the extent to which any number of varieties are similar in terms of the probabilistic grammars that regulate any number of variables and alternations. The nine World Englishes included in our case study are overall remarkably similar to each other in terms of variation patterns: on a scale from 0 (total dissimilarity) to 1 (total similarity), core grammar scores range between $\Gamma = 0.7$ and $\Gamma = 0.8$, which is another way of saying that there is overall strong overlap with regard to the probabilistic grammars regulating variation. In other words, we are dealing with a rather solid "common core" (Quirk et al., 1985, p. 33) of the grammar of English. However, all grammatical alternations are not equal: we saw that the genitive alternation tends to be more stable across varieties than the other alternations. We interpret this as indicating that the alternations under study are differentially sensitive to "probabilistic indigenization," which Szmrecsanyi et al. (2016, p. 133) define as "as the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties." Szmrecsanyi et al. (2016, p. 133) further speculate that "the more tightly associated a given syntactic alternation is with concrete instantiations involving specific lexical items [. . .] the more likely it is to exhibit cross-varietal indigenization effects." Note now that the genitive alternation is an almost entirely abstract alternation without lexical anchors, unlike the dative and—in particular—the particle placement alternation.

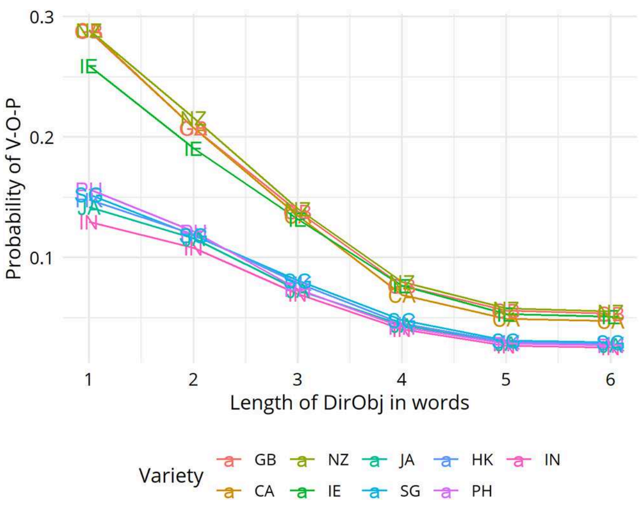


FIGURE 6 | Random forest partial dependence plots of the interaction of variety with direct object length.

Experimentation with subsets of the datasets further showed that spoken language production tends to be more heterogeneous and regionally unstable than written language production (that is, similarity coefficients are lower when attention is restricted to spoken materials). This may be surprising to all those who would like to emphasize that the production of spoken language is subject to processing and production constraints and biases (Hawkins, 1994; MacDonald, 2013) in a way that the production of written language is probably not. But then again, it is a well-known fact that while especially vernacular speech is "the style in which the minimum attention is given to the monitoring of speech" (Labov, 1972, p. 208), written language is more "governed by prescription" (D'Arcy and Tagliamonte, 2015, p. 255), a fact that may level out regional differences. We also saw that Inner Circle varieties form a tighter typological cluster (i.e., similarity coefficients are higher) than the Outer Circle varieties, where similarity coefficients are lower. We speculate that the comparative heterogeneity of Outer Circle varieties is likely due to substrate and contact influences, which play a more important role in the Outer Circle than in the Inner Circle.

In section Mapping Out (dis)Similarity Relationships Between Varieties we moved on to show how the VADIS method can be used to "map out," as it were, relationships between varieties, using techniques and visualization methods (in this case Multidimensional Scaling) widely used in dialectometry and quantitative typology. For the dative alternation, no clear picture emerged, but the plots for the genitive alternation and the particle placement alternation indicated that the Inner Circle varieties tend to cluster together. This is a pattern that has also been reported in the dialect-typological literature based on the aggregate analysis of survey data (see, e.g., Szmrecsanyi and Kortmann, 2009; **Figure 2**). Let us discuss the underlying variation patterns that VADIS is picking up here in more detail. As far as the genitive alternation is concerned, we know, for instance, that Inner Circle users are more sensitive to the *s*-genitive-favoring effect of possessor animacy than Outer Circle users (Heller et al., 2017, p. 18). In regard to the particle placement alternation, the dataset analyzed in Grafmiller and Szmrecsanyi (2018), Grafmiller (2018) shows that users of Inner Circle varieties are more sensitive to the length of the direct object than users of Outer Circle varieties. Consider **Figure 6**, which shows how across all varieties under study, the probability of the split variant (as in *I looked the word up*) decreases as the length of the direct object increases. This is the expected relationship as per the principle of end weight (Behaghel, 1909; Arnold et al., 2000). Note however how the relationship is weaker for the Outer Circle varieties (blueish lines) than for the Inner Circle varieties (yellowish lines). In other words, the principle of end weight is a more potent probabilistic predictor in Inner Circle varieties than in Outer Circle varieties. It is precisely probabilistic contrasts like these that VADIS is designed to be sensitive to.

Next we explored in section Assessing Coherence the extent to which there is coherence between (a) different lines of evidence and (b) between alternations. As to coherence between the different lines of evidence, our data suggest that there tends to be overlap between the 1st line of evidence (constraint significance) and the 2nd line of evidence (effect size), as well as between the 2nd line of evidence and the 3rd line of evidence (constraint ranking). This is true for the genitive alternation and the particle placement alternation; the distance matrices generated on the basis of data from the dative alternation do not overlap at all. As to coherence between alternations, here again the dative alternation is an outlier: the distance matrices derived from the genitive and particle placement alternations do overlap substantially, but the dative alternation distance matrix does not overlap with any of the other distance matrices. The deeper theoretical question that we are addressing here is whether grammar (or the variable parts of grammar) is essentially a collection of independent and/or independently conditioned alternations, or whether alternations actually "agree," as it were, about differences between varieties. Our analysis suggest that we are dealing with a mixed picture. It is unexpected that and unclear why the dative alternation does not pattern with the other alternations: all three alternations are, after all, syntactic/positional alternations that are constrained by similar factors (constituent length, animacy, and so on). Further work is needed to elucidate why the dative alternation is different from the other alternations. It may be worth considering in this connection Guy (2013), a study that investigates if people consistently use stigmatized or prestige variants. Guy finds that it is not easy to demonstrate correlations in the behavior of variables, even if they are generally thought to vary along the same social dimension. The methodology in Guy (2013) is not quite comparable to ours, and he is primarily interested in social variation, not regional variation; but still, the tenor of this work is fully relevant:

> every speech community has many sociolinguistic variables, do the multiple variables cohere in forming sociolects? Thus if each variable has a variant considered 'working class', do working class speakers use all such variants simultaneously? Lectal coherence would imply that variables are correlated; if they are not, the cognitive and social reality of the "sociolect" is problematic (p. 63).

Against this backdrop, the fact that alternations do not cohere perfectly calls into question maybe not so much the reality of World Englishes but conceptions of grammar that consider grammar the aggregation of binary alternations.

One limitation of the VADIS method is that it has many free parameters—in terms of, e.g., the number of constraints to be included in the analysis, regression model structure (random intercepts, slopes, the number of constraints), methods to calculate distance matrices, and so on. This paper has suggested a number of reasonable default parameter settings to address this issue. However, we stress that decisions regarding model parameters, e.g., random effects structure, interactions, and non-linear terms in regression models or the number of trees in the random forests, are best left to individual researchers to determine based on the theoretical questions of interest, as well as the size and composition of their particular datasets. Given the risks of compounding potential problems across multiple models, careful consideration of appropriate model structures and (hyper)parameters is therefore a crucial first step in the analysis. But this step is one that must be evaluated on a case-by-case basis.

Additionally, it is worth reiterating that the validity and reliability of the VADIS method depends upon the quality and representativeness of the data sources. The present study compares standard national varieties at the most general level, and we chose the best available corpora (ICE and GloWbE) for this task. But these sources are not without their issues. Despite the best efforts of ICE compilation teams, social and demographic information is not available for some speakers, and the sampling, and hence representativeness, of some registers in each component will vary somewhat depending on the availability of English texts/speakers in a given region. GloWbE, a massive, aggregate corpus of online texts from around the world, has also been criticized for the unknown degree of variability and heterogeneity in its data sources (see e.g., Davies and Fuchs, 2015 and responses in the same issue). We therefore add a word of caution about generalizing too far beyond the present study, and stress the need for more focused comparisons of individual registers and/or regions.

On a related note, a further aspect that needs to be addressed in future work is external validation of the VADIS methodology. This paper has presented just a first case study showcasing the method and its potential, but comparing the outcome of VADIS to other types of data will be primordial to fully assess the method's strengths and limitations. We are currently exploring ways to use experimental data on speaker intuitions about the three alternations studied in this paper to provide a first step toward external validation of VADIS. Another way to externally validate the outcome of VADIS would be to use correlation analysis to determine how well the distance matrices obtained in VADIS' three lines of evidence align with distance matrices derived from other data on the alternations under study. An example of how this can be done in future work can be found in Röthlisberger (2018) who compares distance matrices derived from probabilistic models to distance matrices calculated based on morphosyntactic information found in the *Electronic World Atlas of Varieties of English* (Kortmann and Lunkenheimer, 2013).

And this takes us to directions for future research, which include the following. The case study presented here is obviously just a first step, and the similarity coefficients and core grammar scores we presented need comparative contextualization. In the realm of English linguistics, we need to include more or different alternations (including phonological, morphological, and function word alternations), and the analysis needs to be extended to more or different regional varieties of English. Beyond English linguistics, we need comparative analysis covering other languages: how stable or unstable are the probabilistic grammars of varieties of e.g., Spanish or French compared to varieties of English? Do we see the same sort of split between native and non-native varieties? And so on. Last but not least, VADIS can be adapted to study not geographical varieties (as we did here) but historical and situational varieties. VADIS could then be used to measure probabilistic stability across time and registers. Recent work in this respect is quite promising. Grafmiller (2018), for example, adopts a VADIS-like approach to investigate stylistic variation in English genitives, and finds that the methods yield patterns in accordance with previous work on register variation. He shows that genitive use in press writing, though still quite distinct from spoken genitives, nevertheless became increasingly more informal/colloquial (e.g., Jucker, 1993) over the twentieth century. Over the same time period, genitives in academic writing also changed dramatically, albeit in ways that do not track with typical colloquialization trends (see e.g., Biber and Gray, 2016; Hyland and Jiang, 2017).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/3gfqn/.

AUTHOR CONTRIBUTIONS

BS, JG, and LR collaborated on the conception and design of the study. Data was collected and prepared by BS and JG. A first draft of the paper was written by BS. JG and LR wrote sections of the paper. BS took care of the final and submitted version of the manuscript which was read, revised, and approved by LR and JG.

ACKNOWLEDGMENTS

Thanks go to all three referees for excellent feedback. Generous financial support from the Research Foundation Flanders (FWO, grant # G.0C59.13N) is gratefully acknowledged.

REFERENCES

Arnold, J. E., Losongco, A., Wasow, T., and Ginstrom, R. (2000). Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28–55. doi: 10.1353/lan.2000.0045

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed effect models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25, 110–142. doi: 10.1515/bgsl.1909.1909.34.530

Biber, D., and Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing (Studies in English Language)*. Cambridge, United Kingdom: Cambridge University Press. doi: 10.1017/CBO9780511920776

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). "Predicting the dative alternation," in *Cognitive Foundations of Interpretation,* eds G. Boume, I. Krämer, and J. Zwarts (Amsterdam: Royal Netherlands Academy of Science), 69–94.

Cysouw, M. (2013). "Disentangling geography from genealogy," in *Space in Language and Linguistics,* eds P. Auer, M. Hilpert, A. Stukenbrock, and B. Szmrecsanyi (Berlin, Boston, MA: Ds Gruyter). Available online at: http://www.degruyter.com/view/books/9783110312027/9783110312027.21/9783110312027.21.xml (accessed January 31, 2015).

D'Arcy, A., and Tagliamonte, S. A. (2015). Not always variable: probing the vernacular grammar. *Lang. Variation Change* 27, 255–285. doi: 10.1017/S0954394515000101

Davies, M., and Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *Engl. World Wide* 36, 1–28. doi: 10.1075/eww.36.1.01dav

De Vaus, D. A. (2002). *Analyzing Social Science Data*. London; Thousand Oaks: SAGE.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* 27, 2865–2873.

Goebl, H. (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

Grafmiller, J. (2018). "Comparative sociolinguistics beyond the vernacular: applying variationist methods to genre variation in written English," *ISLE* 5.

Grafmiller, J. (2018). "When context shapes grammar: stylistic flexibility in the English genitive alternation" in *Presented at the International Congress of Linguists 20* (Cape Town).

Grafmiller, J., and Szmrecsanyi, B. (2018). Mapping out particle placement in Englishes around the world: a study in comparative sociolinguistic analysis. *Lang. Variation Change* 30, 385–412. doi: 10.1017/S0954394518000170

Greenbaum, S. (1991). ICE: the International Corpus of English. *Engl. Today* 7:3. doi: 10.1017/S0266078400005836

Gries, S. T. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York, NY: Continuum Press.

Grieve, J. (2016). *Regional Variation in Written American English (Studies in English Language)*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139506137

Guy, G. R. (2013). The cognitive coherence of sociolects: how do speakers handle multiple sociolinguistic variables? *J. Pragmatics* 52. 63–71. doi: 10.1016/j.pragma.2012.12.019

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge, NY: Cambridge University Press.

Heller, B. (2018). *Stability and fluidity in syntactic variation world-wide: the genitive alternation across varieties of English* (PhD dissertation). KU Leuven, Leuven.

Heller, B., Szmrecsanyi, B., and Grafmiller, J. (2017). Stability and fluidity in syntactic variation world-wide: the genitive alternation across varieties of English. *J. Engl. Linguist.* 45, 3–27. doi: 10.1177/0075424216685405

Hyland, K., and Jiang, F. K. (2017). Is academic writing becoming more informal? *Engl. Spec. Purposes* 45, 40–51. doi: 10.1016/j.esp.2016.09.001

Jucker, A. (1993). "The genitive versus the of-construction in newspaper language," in *The Noun Phrase in English. Its Structure and Variability,* ed A. Jucker (Heidelberg: Carl Winter), 121–136.

Kachru, B. B. (1985). "Standards, codification and sociolinguistic realism: the English language in the outer circle," in *English in the World: Teaching and Learning the Language and Literatures*, eds R. Quirk and H. G. Widdowson (Cambridge: Cambridge University Press), 11–30.

Kachru, B. B. (eds.). (1992). *The Other tongue: English across cultures (English in the Global Context). 2nd Edn*. Urbana: University of Illinois Press.

Kortmann, B., and Lunkenheimer, K. (eds.). (2013). *eWAVE. Leipzig: Max Planck Institute for Evolutionary Anthropology*. Available online at: http://ewave-atlas.org/ (accessed 31 July, 2019).

Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling*. Newbury Park, London, New Delhi: Sage Publications. doi: 10.4135/9781412985130

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762. doi: 10.2307/412333

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Philadelphia Press.

Leimgruber, J. R. E. (2013). *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139225755

Levshina, N. (2015). *How to Do Linguistics With R: Data Exploration and Statistical Analysis*. Amsterdam; Philadelphia, PA: John Benjamins Publishing Company. doi: 10.1075/z.195

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Front. Psychol.* 4:226. doi: 10.3389/fpsyg.2013.00226

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. *J. Mem. Lang.* 94, 305–315. doi: 10.1016/j.jml.2017.01.001

McArthur, T. (1998). *The English Languages*. Cambridge: Cambridge University Press. doi: 10.1017/9780511621048

Nerbonne, J., Heeringa, W., and Kleiweg, P. (1999). "Edit distance and dialect proximity," in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison,* Vol. V–XV. eds D. Sankoff and J. Kruskal (Stanford, CA: CSLI Press).

Poplack, S., and Tagliamonte, S. (2001). *African American English in the Diaspora (Language in Society 30)*. Malden, MA: Blackwell.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London; New York, NY: Longman.

Rosenbach, A. (2008). Animacy and grammatical variation: findings from English genitive alternation. *Lingua* 118, 151–171. doi: 10.1016/j.lingua.2007.02.002

Röthlisberger, M. (2018). *Regional variation in probabilistic grammars: a multifactorial study of the English dative alternation* (Ph.D. dissertation). KU Leuven, Leuven. Available online at: https://lirias.kuleuven.be/handle/123456789/602938

Röthlisberger, M., Grafmiller, J., and Szmrecsanyi, B. (2017). Cognitive indigenization effects in the English dative alternation. *Cogn. Linguist.* 28, 673–710. doi: 10.1515/cog-2016-0051

Schneider, E. (2011). *English Around the World: An Introduction*. [S.l.]: Cambridge University Press.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35, 335–357.

Spruit, M. R., Heeringa, W., and Nerbonne, J. (2009). Associations among linguistic levels. *Lingua* 119, 1624–1642. doi: 10.1016/j.lingua.2009.02.001

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14, 323–348. doi: 10.1037/a0016973

Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry (Studies in English Language)*. Cambridge, New York, NY: Cambridge University Press.

Szmrecsanyi, B., Grafmiller, J., Heller, B., and Röthlisberger, M. (2016). Around the world in three alternations: modeling syntactic variation in varieties of English. *Engl. World Wide* 37, 109–137. doi: 10.1075/eww.37.2.01szm

Szmrecsanyi, B., and Kortmann, B. (2009). The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119, 1643–1663. doi: 10.1016/j.lingua.2007.09.016

Szmrecsanyi, B., and Röthlisberger, M. (2019). "World Englishes from the perspective of dialect typology," in *The Cambridge Handbook of World Englishes*, eds M. Hundt, E. W. Schneider, and D, Schreier (Cambridge: Cambridge University Press).

Tagliamonte, S. (2001). "Comparative sociolinguistics," in *Handbook of Language Variation and Change,* eds J. Chambers, P. Trudgill, and N. Schilling-Estes (Malden; Oxford: Blackwell), 729–763. doi: 10.1002/9780470756591.ch28

Tagliamonte, S. (2006). *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tagliamonte, S. (2012). *Variationist Sociolinguistics Change, Observation, Interpretation*. Malden, MA: Wiley-Blackwell. Available online at: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=819316 (accessed 31 July, 2019).

Tagliamonte, S. A., D'Arcy, A., and Louro, C. R. (2016). Outliers, impact, and rationalization in linguistic change. *Language* 92, 824–849. doi: 10.1353/lan.2016.0074

Tamaredo, I., Röthlisberger, M., Grafmiller, J., and Heller, B. (2019). Probabilistic indigenization effects at the lexis–syntax interface. *Engl. Lang. Linguist.* doi: 10.1017/S1360674319000133. [Epub ahead of print].

Tamminga, M., MacKenzie, L., and Embick, D. (2016). The dynamics of variation in individuals. *Linguis. Variation* 16, 300–336. doi: 10.1075/lv.16.2.06tam

van Hout, R., and Muysken, P. (2016). "Taming Chaos. Chance and variability in the language sciences," in *The Challenge of Chance*, eds K. Landsman and E. van Wolde (Cham: Springer International Publishing), 249–266. http://link.springer.com/10.1007/978-3-319-26300-7_14 (accessed September 21, 2018).

Wasow, T., and Arnold, J. (2003). "Post-verbal constituent ordering in English," in *Determinants of Grammatical Variation in English,* eds G. Rohdenburg and B. Mondorf (Amsterdam: Mouton de Gruyter), 119–154. doi: 10.1515/9783110900019.119

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 09 April 2020
doi: 10.3389/frai.2020.00015

Hybrid Hashtags: #YouKnowYoureAKiwiWhen Your Tweet Contains Māori and English

David Trye[1], *Andreea S. Calude*[2]*, *Felipe Bravo-Marquez*[3] *and Te Taka Keegan*[1]

[1] Department of Computer Science, University of Waikato, Hamilton, New Zealand, [2] School of General and Applied Linguistics, University of Waikato, Hamilton, New Zealand, [3] Department of Computer Science, University of Chile & IMFD, Santiago, Chile

Twitter constitutes a rich resource for investigating language contact phenomena. In this paper, we report findings from the analysis of a large-scale diachronic corpus of over one million tweets, containing loanwords from te reo Māori, the indigenous language spoken in New Zealand, into (primarily, New Zealand) English. Our analysis focuses on hashtags comprising mixed-language resources (which we term *hybrid hashtags*), bringing together descriptive linguistic tools (investigating length, word class, and semantic domains of the hashtags) and quantitative methods (Random Forests and regression analysis). Our work has implications for language change and the study of loanwords (we argue that hybrid hashtags can be linked to loanword entrenchment), and for the study of language on social media (we challenge proposals of hashtags as "words," and show that hashtags have a dual discourse role: a micro-function within the immediate linguistic context in which they occur and a macro-function within the tweet as a whole).

Keywords: language contact, loanwords, hashtags, hashtag half-life, Māori, New Zealand English, word embeddings, the language of social media

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Rolando Alberto Coto-Solano,
Dartmouth College, United States
Kristina Kocijan,
University of Zagreb, Croatia

****Correspondence:***
Andreea S. Calude
andreea.calude@waikato.ac.nz

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *19 December 2019*
Accepted: *13 March 2020*
Published: *09 April 2020*

Citation:
Trye D, Calude AS, Bravo-Marquez F and Keegan TT (2020) Hybrid Hashtags: #YouKnowYoureAKiwiWhen Your Tweet Contains Māori and English. Front. Artif. Intell. 3:15. doi: 10.3389/frai.2020.00015

1. INTRODUCTION

Languages, like people, rarely exist in complete isolation from one another. One of the most predictable outcomes of language contact, brought about by contact between speakers of (distinct) languages or language varieties, is the adoption of new words from one language (variety) into another. Languages are "leaky" (parallel to Sapir, 2004, p. 29) and speakers act like fluid transmitters of words between the languages they navigate. While linguists have studied loanwords for decades (see work dating back to the 1950s, e.g., Haugen, 1950; Weinrich, 1953), the fruits of this labor can be roughly summarized in three main strands, all of which focus primarily on the borrowing process as a linguistic matter: (1) studies focusing on what is (or can be) borrowed (e.g., Field, 2002; Haspelmath and Tadmor, 2009; Matras, 2009; inter alia), (2) studies attempting to distinguish (if possible) between loanword use and code-switching (e.g., Muysken, 2000; Stammers and Deuchar, 2012; Backus, 2013 and others), and (3) studies which document the adaptation of the loaned material to the internal rules of the receiver language, whether phonological or morphological (e.g., Poplack and Sankoff, 1984; Poplack et al., 1988; Hashimoto, 2019 and references cited within).

In recent decades, a paradigm shift has unfolded in the study of loanwords, which considers linguistic borrowing in its wider sociolinguistic context. In this view, borrowing is not just a linguistic event but also a socially meaningful one, placing both language and speaker at its center. The "socio-pragmatic turn" of loanword study, discussed in a recent *Special Issue* on the topic by Zenner et al. (2019), is shifting to include matters beyond language prestige, such as identity, language ideology, and cultural knowledge (captured by the term "language regard"; see Preston, 2013). Our study seeks to complement this body of work by bringing in the dimension of *language play*. We show that the loanwords in our data are used creatively to signal solidarity with and belonging to an indigenous group, which, despite being previously marginalized, is gaining visibility and status in the wider community. The social dimension of the loanwords we discuss here is undeniably strong and it is virtually impossible to make sense of the borrowing process in this case without recourse to the aforementioned notion of language regard.

The current study examines an unusual language contact situation, as described below. We report findings from an empirically-driven, corpus linguistics analysis of Māori loanwords in (primarily) New Zealand English (NZE) by exploring a purpose-built, large-scale dataset of social media language from the Twitter platform. Examples (1–3)[1] illustrate the phenomenon in question (loanwords are given in bold text):

> (1) Sorry I thought you were **Kiwi** [a New Zealander]. **Aotearoa** is the **Māori** name for NZ [ID 1064121983678406656]
>
> (2) We stand united Native American **Whanau** [family], **kia kaha** [be strong] DakotaAccessPipeline **#haka** [war dance] **#Maori** **#whanau** #NativeAmerican #united [ID 793003612217577472]
>
> (3) I'm **Pākehā** [European New Zealander] and went to a majority **Māori** primary school. There was lots of incorporation of **#tereo** [the Māori language] and **tikanga** [customs] into everyday activities, set me on path to wanting to live in bicultural **aotearoa** #letssharegood**tereo**stories [ID 959155122289823744]

The language contact situation between the indigenous Austronesian language of te reo Māori and (New Zealand) English presents a unique opportunity to study the flow of words from an endangered, minority-status language (te reo Māori) into a dominant, global *lingua franca* (English). The direction of lexical transfer, especially on the scale of that observed in New Zealand English is, to our knowledge, not comparable to any other language situation previously described (for a detailed description of the nature of the contact situation between Māori and English in New Zealand, see section 3 in Levendis and Calude, 2019 and section 3.1 in Calude et al., 2017).

The study of Māori loanwords in New Zealand English has received intense scrutiny in the literature, especially with regard to newspaper language (Davies and MacLagan, 2006; Macalister, 2006, 2009; Onysko and Calude, 2013), Hansard Parliament debates (Macalister, 2006), children's picture books (Daly, 2007, 2016), TV language (de Bres, 2006), conversation (Kennedy and Yamazaki, 1999), and more recently, online science discourse (Calude et al., 2019b). However, very little is known about the use of Māori loanwords on social media (with the exception of a small sample of tweets in Calude et al., 2019b, and preliminary findings in Trye et al., 2019), which motivates our attention to Twitter data here.

The large body of work cited above has uncovered a number of trends regarding the use of loanwords in New Zealand English. Perhaps the most important one relates to their diachronic use, which strongly suggests that their use is increasing over time (Kennedy and Yamazaki, 1999; Macalister, 2006; Calude et al., 2019a). Moreover, while European settlers initially borrowed flora and fauna words to refer to the new species they encountered upon arriving in New Zealand (e.g., *kiwi*, *rimu*, and *kauri*), over time, as the new variety of English began to emerge, it started to adopt more material and social culture words (e.g., *marae*, *tangi*, and *powhiri*; see Macalister, 2006). Secondly, the use of Māori loanwords is driven by Māori women and is largely associated with Māori-related discourse topics (Kennedy and Yamazaki, 1999; de Bres, 2006; Degani, 2010; Calude et al., 2017). Calude et al. (2017) further found that certain loanwords appear to be "more successful" compared to others. Loanword success is defined as the chance of a loanword being used within a receiving language, compared to an existing lexical alternative word native to the receiving language, controlling for the number of opportunities that speakers of the receiving language have to use the concept denoted by the loanword. For instance, loanwords which are shorter than their native English counterpart (in terms of number of syllables, e.g., *pā*/settlement, *tangi*/funeral, *reo*/language) are comparatively more successful, as well as loanwords that encode cultural rather than core meanings (in the sense of Myers-Scotton, 2002). The study also found that linguistic factors interacted with the sociolinguistics ones, such that, for Māori speakers, the ethnicity of the audience had a role to play (when speaking to a Māori-only group, Māori speakers seemed more sensitive to efficiency effects), and, for Pākehā (European) New Zealanders, polysemous loanwords were comparatively less successful than monosemous loanwords (ibid).

In light of what is currently known about Māori loanwords in New Zealand English, we wanted to investigate their use on social media. To this end, we investigated data from Twitter—in part, due to practical considerations (the ease of collecting electronically-searchable data), and in part because this data complements the other types of genres previously investigated. Like spoken, conversational language, Twitter language is (largely) informal, unplanned, non-editable, and immediately available to potential audiences and, like newspaper language, Twitter language is written down. Furthermore, Twitter users span both ends of the formal spectrum, from individuals reflecting their own linguistic style (with regard to lexical content, spelling, word play, etc.) to institutions representing collectives of various sizes (Universities, political parties, etc.) who are perhaps more likely to conform to social norms. However, collecting a corpus of Twitter language for our specific purposes, namely, studying Māori loanwords in New Zealand English, is not without its problems, as discussed in section 3.

[1] To the best of our knowledge, the examples of tweets we include in this paper comply with the terms and conditions specified by Twitter for research use, see https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html.

One of the most distinctive uses of Māori loanwords in our Twitter corpus, once collected, was the use of *hybrid hashtags*. These are hashtags which involve (at least) one word of Māori and (at least) one word of (NZ) English[2]. Examples include #letssharegoodtereostories (as illustrated in example 3), #kiwigold, #honeyhui, #TreatyofWaitangi, and #beingmaori. We are not aware of any other research that analyses hybrid hashtags specifically, although they are mentioned in passing by Lee and Chau (2018) in their analysis of hashtags on Instagram containing a mixture of English and Cantonese (p. 26). The study of minority languages in social media through hashtag use is not new in itself (see for instance McMonagle et al., 2019), but our focus on combinations of lexical resources from a minority and a majority language in a single hashtag (as opposed to the use of distinct hashtags from different languages in one tweet, as analyzed by Jurgens et al., 2014) has to our knowledge not been studied before. For this reason, the current paper focuses exclusively on the findings uncovered in relation to hybrid hashtags. Before turning our attention to how we built our Twitter corpus and what we found in the data, we first summarize two of the main strands of research questions addressed by recent work on the linguistics of hashtags, in section 2.

2. THE LINGUISTICS OF HASHTAGS

Linguistic analyses of Twitter and social media discourse are becoming increasingly prevalent as the genre captures the attention of language researchers. One feature which started out on social media, but which is already making its way into other genres (see Caleffi, 2015; Evans, 2015) is the hashtag. Hashtags (denoted with a "#" symbol) have been described as a means of "[categorizing] messages posted on Twitter" (Cunha et al., 2011, p. 58), or of "referring to a topic and creating communities of people interested in that topic" (Caleffi, 2015, p. 67). Adopting a discourse-based approach, Page (2012) conceptualizes Twitter as a "linguistic marketplace," in which hashtags are a crucial currency. Zappavigna (2011) argues that hashtags function as a "community building linguistic activity" (p. 789) that enables "ambient affiliation" (p. 790).

However, even in this very much emerging body of work, two main preoccupations stand out. First, there are surging debates about the morphological processes which give rise to hashtags. Two main arguments have been proposed so far, which might be succinctly summarized as "hashtags as compounds" (Maity et al., 2016) and "hashtags as hashtagging" (Caleffi, 2015). However, the evidence is still moot with regard to these positions. We return to the word-formation process in section 5.1.

The second open question that has generated interest in the hashtag literature relates to what influences the life-cycle of a hashtag. Given that hashtags are essentially a new brand of "word," even if only comprising an existing, single word (e.g., #fun), the fact that the word is used together with the "#" symbol and functions as a hashtag distinguishes it both orthographically, semantically and functionally from its use without the "#" symbol. This lexical (re-)birth constitutes a linguistic innovation which means that the hashtag, like all other members of the lexicon of a language, has to "fight for its survival" in order to avoid falling out of use. Romero et al. coin two terms in relation to hashtag life-cycle, namely *persistence*—"the extent to which repeated exposures to a hashtag continue to have a marginal effect" (Romero et al., 2011, p. 695) and *stickiness*—"the probability of adoption based on one or more exposures" (ibid). The term *persistence* is problematic because exposure refers, in practice, to frequency of use of a hashtag, but not necessarily to its likelihood of being seen by other Twitter users (as the word "exposure" suggests), because users do not necessarily read all posts written by users in their Twitter network. *Stickiness* is similarly problematic because of the assumptions encapsulated by the word "exposure." However, it is certainly possible to use frequency of use of various hashtags on Twitter as a measure of hashtag survival in this genre, assuming that the longer a hashtag is used, the longer its lifespan, life-cycle or survival[3].

In this paper, we propose (what we believe to be) a more informative measure of a hashtag's success, namely, a hashtag's "half-life," based on the concept of a word's half-life, introduced by Pagel and Meade (2006). Pagel and Meade define a word's half-life as the amount of time by which a given word has a 50% chance of being replaced by a non-related (non-cognate) form (Pagel and Meade, 2006, 2018; Pagel et al., 2007). By analogy, our notion of a hashtag's half-life refers to the amount of time by which a hashtag reaches half of its total impact (or activity), where "impact" is measured in total number of uses (that is, a frequency of use measure). We return to this in section 4.2.

Regardless of our evaluation of the notions of persistence and stickiness, the most important finding from Romero et al. (2011) in relation to longevity of hashtags pertains to the semantic domain of the various hashtags investigated: hashtags from controversial political topics appear to be more sticky and persistent, whereas hashtags encoding idioms are comparatively less sticky and persistent (2011, p. 701). This finding has informed our own work and we look to the semantic domain of the various hashtags we analyze in relation to hashtag success.

Other studies have also tried to model hashtag longevity by considering various factors. Cunha et al. (2011, pp. 63-64) found an inverse relationship between a hashtag's length and its longevity, and a decrease in longevity associated with the use of underscores in hashtags. Maity et al. (2016, p. 60) investigated two-word compound hashtags (#AB, where A and B are free morphemes) and found that "propagation" of such hashtags is most significantly correlated with an increase in overlap of the lexical content of tweets containing the single-word hashtags (i.e., #A and #B). Tsur and Rappoport (2012) investigate four types of features in relation to hashtag popularity: (1) features concerning the linguistics of the hashtag itself, such as length, position in the tweet, and others, (2) features concerning the content of the tweet containing the hashtag investigated (e.g., tweet length), (3) features to do with the user data of the tweet containing the hashtag in question (e.g., number of followers), and (4) features to do with the temporal patterns of use of

[2] In our data, we also included #hakarena, which comprises one morpheme (*-rena*, from *Macarena*) and one free word (*haka*).

[3] We use these terms interchangeably.

the hashtag (normalized weekly counts). They tested these four features as a "bundle" (not separately) and found that, of the four feature types, hashtag content features and tweet content features contributed only a marginal increase in the prediction of hashtag popularity (although they did seem to contribute toward reduced error rates, see p. 649 ff.). The features that do best with regard to predicting hashtag popularity are features to do with user data and timestamps.

3. MATERIALS AND METHODS

This section documents our corpus and the methods we used to build it. We first discuss the Twitter corpus and provide an overview of how we created it, and then focus our attention on the data containing the hybrid hashtags and the sub-corpus we extracted to study these.

3.1. Building the Māori Loanword Twitter (MLT) Corpus

The *Māori Loanword Twitter (MLT) Corpus*[4] was created using a novel technique that relies on a set of query words, instead of following specific users (cf. Keegan et al., 2015) or tracking geolocations (cf. Grieve et al., 2017). This process is briefly summarized below, but a more detailed explanation is given in Trye et al. (2019).

First, we used the Twitter Search API[5] to obtain 8 million tweets containing one or more query words. The vast majority of these words were compiled by Hay (2018), as part of a study identifying Māori words that most monolingual, English-speaking New Zealanders recognized, even if they did not know their meaning (for the full list of query words, see **Table S1**). Given the high level of recognition associated with these words, we predicted that they were likely to be used in New Zealand English tweets, and as such, would make a suitable starting point for building the corpus.

However, inspection of the data revealed that many query words frequently occurred in non-New Zealand English contexts, and some were seldom used as loanwords (particularly short, three- or four-letter words with multiple meanings in different languages). We addressed this noise by using supervised machine learning, the problem being analogous to spam classification (see Abayomi-Alli et al., 2019). After manually labeling a sample of tweets for each query word as "relevant" or "irrelevant," we removed tweets containing query words that were irrelevant more than 90% of the time and trained a classifier to automatically determine when the remaining query words were used in relevant (New Zealand English) contexts. In this way, we could filter out irrelevant tweets to produce a higher-quality corpus.

Drawing on lessons learned from the original study (Trye et al., 2019), some improvements were made to further mitigate noise in the MLT corpus. First, the corpus was enhanced by deploying a Multinomial Naive Bayes model (McCallum and Nigam, 1998) that considered not only unigrams in the feature space (as per the previous study), but bigrams as well. Using the same stratified training set as before, superior Kappa and F-score values were achieved (0.5754 and 0.819, respectively), along with a matching AUC value of 0.872. Additionally, following the removal of tweets classified as irrelevant by the model, 81,830 duplicate tweets were discarded. These duplicates were the result of some tweets containing multiple query words, and being harvested independently by each occurrence.

TABLE 1 | Corpus statistics for the MLT corpus, by year.

Year	Tweets	Words	Users
2006	8	135	7
2007 △	819	12,872	468
2008 △	5,903	96,665	3,551
2009 △	67,834	1,141,748	38,908
2010 △	142,509	2,310,289	76,713
2011 △	306,389	4,760,881	167,471
2012 △	427,428	6,296,131	241,584
2013 △	446,505	6,630,105	249,388
2014 ▽	345,150	5,254,932	190,181
2015 ▽	315,128	4,847,984	177,482
2016 ▽	240,793	3,741,744	132,867
2017 △	288,779	4,870,311	141,049
2018 △	292,966	6,863,834	143,607
Total	2,880,211	46,827,631	1,226,109

For the (distinct) Users column, there is some overlap across years, because the same users may be active over multiple years (hence the number of distinct users per year does not match the total in the bottom row).

The final MLT corpus consists of 2,880,211 tweets, comprising 46,827,631 word tokens. In total, these tweets capture linguistic output from 1,226,109 distinct users. A diachronic overview is provided in **Table 1**.

3.2. Building the Hybrid Hashtag Sub-corpus

Once collected, we analyzed the MLT corpus for hashtag use. In total, our corpus contains 8,753 distinct hashtags that occur ten times or more (this figure considers alternative spelling, capitalization and punctuation, e.g., macron use, as giving rise to distinct hashtags; therefore, #kiwias and #KiwiAs are counted as separate hashtags).

We manually scanned these hashtags for the presence of Māori and English lexical items, and extracted 287 hashtags that were hybrid. We then discarded hybrid hashtags whose meanings were unclear, even after carefully inspecting the tweets in which they were used (e.g., #kiwifollowspree). Furthermore, we removed hashtags whose meanings were tied to a particular in-group and therefore limited from wider use (e.g., #kiwiPyCon, which refers to a New Zealand-based conference for Python programmers), as well as hashtags denoting specific organizations (e.g., #manaparty), brands (e.g.,

[4] The corpus is available to download at https://kiwiwords.cms.waikato.ac.nz/corpus/

[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

TABLE 2 | Linguistic characteristics of the Māori loanwords used in hybrid hashtags.

Loanword	English counterpart(s)	Semantic category	Core/cultural distinction
Kiwi(s)	Kiwi fruit, flightless bird or New Zealander(s)	Flora & fauna/social culture	cultural
Māori	(Of) indigenous (origin)	Social culture	cultural
haka	Tribal dance	social culture	cultural
(te) reo	Pertaining to Maori language or to (any) language	Social culture	core
hui	Meeting	Social culture	core
Waitangi	Place name	Proper noun	cultural
Aotearoa	New Zealand	Proper noun	cultural
kai	Food	Material culture	core
kia ora	Hello, thank you, goodbye	Social culture	cultural

The loanwords are given in order of raw frequency in the HH sub-corpus from most to least frequent. We follow Macalister (2006) for semantic categories of loanwords and Myers-Scotton (2002) for core/cultural distinctions.

#maoritv), and sports teams (e.g., #KiwiFerns, used for New Zealand Rugby League).

We primarily wanted to discard hybrid hashtags that were proper nouns because, by and large, these hashtags did not constitute a meaningful linguistic choice (for example, #voteMarama, where "Marama" is the name of a person). However, we did retain six hashtags that were proper nouns, because we wanted to compare their use with content noun phrases and hashtags functioning as other word classes (verbs, clauses, etc.). Of the six proper-noun hashtags, three denote various ethnic or national groups (#MeanMaori, #AotearoaNZ, and #NZMaori), two denote regularly occurring, large-scale, national events (#WaitangiDay[6] and #MaoriLanguageWeek[7]) and the last hashtag, #TreatyofWaitangi, denotes the most defining event in New Zealand history.

This process whittled down our list of hybrid hashtags from 287 to 135 hashtags. Since the remaining hashtags contained variations in capitalization, macron use, and inflections, we amalgamated them into 81 hybrid hashtag lemmas (e.g., #gokiwis, #goKiwi, and #GOKIWIS were all coded under the single hybrid #gokiwi(s) in our data, and #beingMāori—with a macron—was combined with #beingMaori—without one). The 81 hybrid hashtags were used in 5,684 tweets in total (from the MLT corpus), and posted to Twitter by 3,771 distinct users. These hashtags and their associated tweets comprise the hybrid hashtag dataset—hereafter, the *HH sub-corpus*[8]. For further details about how this corpus was created, please see **Supplementary Material**, Section 1.

4. RESULTS

This section outlines the results of the 81 hybrid hashtags analyzed in the HH sub-corpus. We begin by outlining general linguistic characteristics of the hashtags, specifically the types of loanwords which occur in the hashtags, and the semantic and syntactic function of the hashtags, as well as their lengths. Section 4.2 discusses measures of hashtag success and predictions of hybrid hashtag success in our corpus.

4.1. General Linguistic Characteristics of Hybrid Hashtags

The first thing to note about the hybrid hashtags in the HH sub-corpus is that the 81 hashtags are created using only nine Māori loanwords. For the most part, these nine loanwords, given in **Table 2**, are documented to be among the top ten most frequent loanwords in other corpora of New Zealand English (for example, the *Wellington Corpus of Spoken New Zealand English*, Holmes et al., 1998; the *Matariki Corpus*, Calude et al., 2019a; and the *Māori Language Week Corpus*, Levendis and Calude, 2019). Secondly, they constitute a mix of core and cultural borrowings (following Myers-Scotton, 2002), with a slight skew toward cultural borrowings. Finally, semantically, they tend to denote social culture terms (following the distinctions proposed by Macalister, 2006).

Among the nine loanwords giving rise to the 81 hybrid hashtags extracted, we find that two loanwords, *kiwi(s)* and *Māori*, are significantly more productive in forming hybrid hashtags than all other loanwords. Overall frequency counts and examples are given in **Table 3**.

Many hybrid hashtags contain semantically positive words (e.g. "loyal," "awesome," "proud," "love," and "good"), which reflect the polarity of the tweet itself. Examples (4) and (5) illustrate this (hybrid hashtags are given in bold text in these and subsequent examples).

> (4) @ClaireLHuxley kiwis impress me anyway but that was over and beyond **#proudkiwi** [ID 123993688413188098]

> (5) I'm proud to have such a strong heritage, my ancestors were warriors **#maoripride** #proud #Maori #aotearoa #whanau #culture [ID 300417134650068992]

Conversely, there is one hybrid hashtag, #BanTheHaka, which is (nearly always) explicitly negative. The haka is a Māori tribal dance that is routinely performed (among other occasions) before international rugby matches, and it is in this capacity that it has gained considerable attention on the world stage. However, the practice has attracted controversy from people who see the behavior as unnecessarily aggressive or intimidating. Example (6) provides an opinion to this effect and example (7) links the haka to an "unfair advantage" to the team performing it. Both these tweets align themselves with the literal and most likely, the original meaning captured by the hashtag #BanTheHaka, which is to express a negative attitude toward the haka.

[6] Waitangi Day is the national day of New Zealand, which takes place in February each year.

[7] Māori Language Week is an annual, government-sponsored initiative to promote Māori language use.

[8] The HH sub-corpus is available to download at https://waikato.github.io/kiwiwords/hh_corpus/

TABLE 3 | Usage statistics for the nine Māori loanwords present in the set of hybrid hashtags.

Loanword	Raw freq.	Hybrid hashtags	Total tweets
kiwi(s)	54	#GoKiwi(s), #proudkiwi(s), #kiwipride, #proudtobe(a)kiwi, #youknowyoure(a)kiwiwhen…	3,487
Māori	12	#beingmaori, #NZMaori, #maorilanguage, #MAORISTYLES, #maoripride…	874
haka	5	#Hakarena, #BanTheHaka, #HakaTime, #thehaka, #lovethehaka	224
(te) reo	3	#LetsShareGoodTeReoStories, #Keep(in)ItReo, #goodtereostories	360
hui	2	#huitweet, #honeyhui	35
Waitangi	2	#WaitangiDay, #TreatyofWaitangi	653
Aotearoa	1	#AotearoaNZ	15
kai	1	#kaitime	15
kia ora	1	#kiaora4that	21
Total	81		5,684

Loanwords are given in decreasing order of raw frequency in the HH sub-corpus. The hybrid hashtags in the third column are listed according to number of tweets, with the most frequently occurring lemma reported for each one. For the loanwords kiwi(s) and Māori there were many more hybrid hashtags than included in the table (only the five most common are shown here; for full details, see ***Supplementary Material****).*

(6) The Haka has never been "Respectful"! It's always been aggressive! **#BANTHEHAKA** [ID 796629023887622144]

(7) @gwladrugby. The Haka is an unfair advantage for NZ to be able to perform b4 the game, should be able to respond how u wish ! **#banthehaka** [ID 128792760386985985]

However, another tweeter in our corpus uses the hashtag to join the discussion surrounding the practice of the haka, but with the aim of presenting the opposite view; namely, writing in support of the tradition.

(8) #BanTheIgnorance instead of ban the haka. Do some research next time you insult an entire culture **#BanTheHaka** [ID 665815361694994432]

These examples illustrate two facets of hashtags. First, hashtags need to be interpreted by examining the global (macro) context within which they are used (here within the entire tweet, not just with reference to the phrase or clause they are part of). Secondly, they can have a dual function within this context of use, one of these functions being the semantic expression of a particular meaning, for instance, in examples (6) and (7), the expression of a negative attitude toward the performance of the haka, and a second function being a discourse affiliative role, namely of contributing to an existing discussion or community of practice (as also argued by Cunha et al., 2011; Caleffi, 2015). Our examples show that the two functions can co-occur without conflict in many tweets [examples (6–7) are such cases], but that it is also possible for the two functions to appear in conflict with each other [as in example (8)], when the literal meaning expressed by the hashtag violates the propositional content of the tweet. In such cases, the conflict is resolved by having the discourse affiliative function override the semantic expression of the hashtag (rendering the hashtag's semantic content moot). We return to these points in the Discussion section.

Given the findings discussed by previous literature on hashtags more generally (see section 2), we also investigated four linguistic properties of our set of hybrid hashtags, including hashtag length and semantic domain (as per previous studies). In addition, we considered whether the hashtags had multiple distinct variables (before amalgamating the lemmas), and looked at each hashtag's syntactic word class[9].

The first linguistic characteristic coded was hashtag length, in number of words (following other work analysing hashtag length, namely Cunha et al., 2011; Tsur and Rappoport, 2012; Maity et al., 2016). **Figure 1**[10] Illustrates the distribution of lengths in the HH sub-corpus (by both number of tweets, **Figure 1A**, and by distinct number of hashtags, **Figure 1B**). As can be seen, these lengths range between two and six words, with most hybrid hashtags consisting of two words.

Next, as discussed in section 3.2, some hashtags had multiple variants (due to slight differences in capitalization, macron use, and/or inflections), whereas others consisted of only one form. For example, the hashtag #flyingkiwis has three variants, which vary in their use of capitals and singular/plural forms: #FlyingKiwis, #flyingkiwis, and #flyingkiwi. As noted above, we did not want to count these hashtags as being distinct so we merged them into the same hashtag lemma. Our corpus of 81 hashtags contains slightly more hashtags with unique forms ($n = 46$) than with multiple variants ($n = 35$). However, the hashtags with multiple variants appear to be used in a higher number of tweets overall (see **Figure S1**).

Third, we consider word-class possibilities for the hybrid hashtags. **Table 4** details the various word-class possibilities realized in our data and provides examples to illustrate these. **Figure 2** shows a frequency distribution of these possibilities in the HH sub-corpus (in terms of number of tweets).

Finally, we turn to the semantic domain of our hybrid hashtags. In accordance with claims by Macalister (2006) for other genres of New Zealand English, we also find that the hybrid hashtags are used to reference New Zealand identity, (NZ) flora and fauna and humor (see also Macalister, 2002), but in addition, we find that they are commonly used in sporting contexts. **Table 5** exemplifies each of the semantic domains uncovered in the HH sub-corpus, and **Figure 3** gives their frequency distribution.

This was by far the hardest linguistic factor to code in our data. Two main sets of problems made the coding difficult. First, some hashtags seemed to belong to multiple semantic categories, either because different tweeters used the hashtag in different ways, or because the same tweeters varied their use of the hashtag (or

[9] We decided to include these factors because the hybrid hashtags in our dataset appear to show considerable variation in regard to both of these.

[10] All figures included are drawn using *R Software* (R Core Team, 2017) and the *ggplot* package (Wickham, 2009).

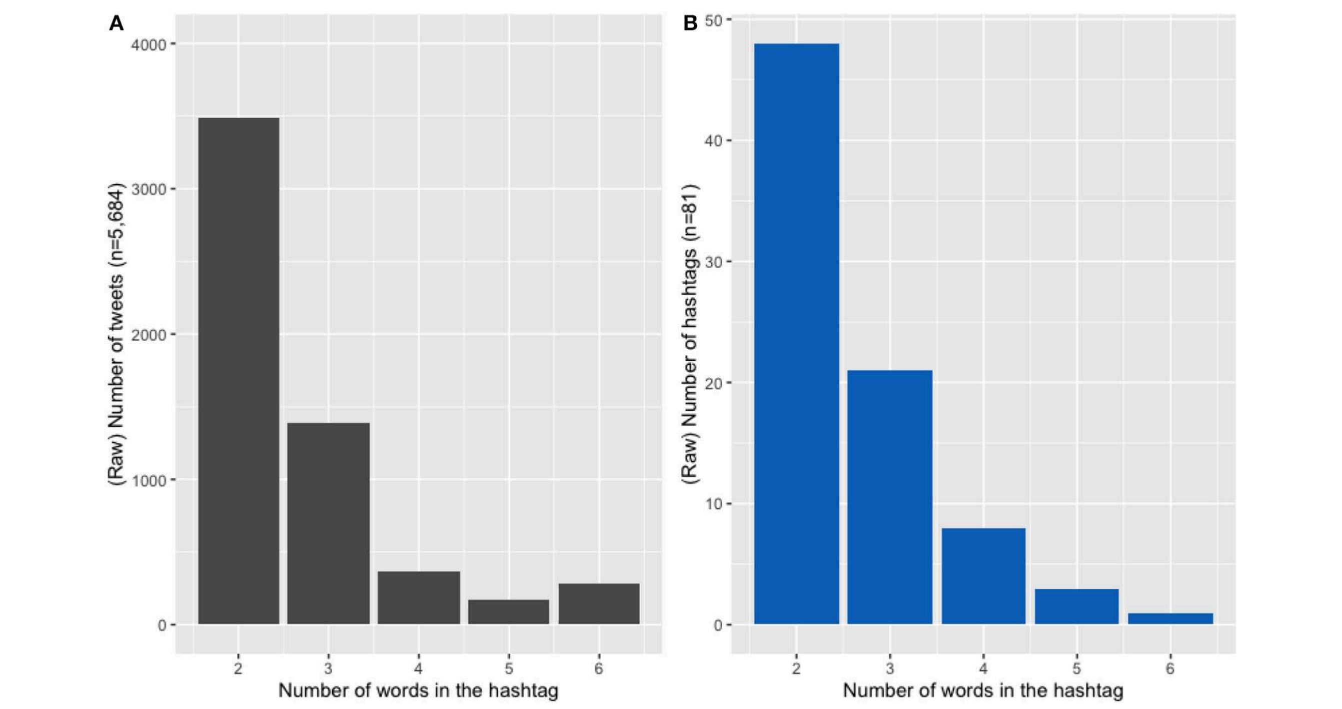


FIGURE 1 | (A) Distribution of hashtag length across number of tweets. **(B)** Distribution of hashtag length by (hashtag) type.

sometimes a combination of both), as shown in examples (9–11). Secondly, the meaning of the hashtag was not always transparent, nor was its use in the tweet. In all cases, we chose the domain that appeared to be the most dominant in the HH sub-corpus (i.e., the domain that applied to the most tweets containing that particular hashtag).

For example, consider the hashtag #kiwiquestion. This hashtag was mostly used by the same tweeter, but sometimes in reference to (native) flora and fauna (9) and sometimes denoting NZ identity (10):

> (9) Here we go, our **#KiwiQuestion** of the day: What are thought to be the kiwi bird's two closest relatives? [ID 288571983359262720]

> (10) **#KiwiQuestion** What do the stars on the New Zealand flag represent? Answer for a #free Shisha from Kiwi. Smokers unite! #Maadi #freestuff [ID 293282293051703297]

Example (11) shows the use of the same hashtag by a different tweeter, in a completely different context (to ask a question about eating kiwifruit, which falls under the "flora and fauna" category):

> (11) Random I know but do you leave the skin on a kiwi fruit when eating it or peel it off? **#kiwiquestion** [ID 177022614559141888]

However, we classified this hashtag as "NZ identity" because most of the tweets were similar to example (10).

In order to alleviate the problems we had in assigning a (single) semantic domain to each hybrid hashtag, we verified our choices by training word embeddings on the MLT corpus and visualizing the semantic neighborhood of the hybrid hashtags in question.

Word embedding algorithms utilize principles of distributional semantics—the notion that similar words occur in similar contexts—to model relationships between words. These algorithms have gained prominence in the field of Natural Language Processing (NLP) in recent years, and are widely regarded as a useful tool for linguistic analysis (when used appropriately). However, word embeddings are not without their limitations, as discussed by Bowern (2019) (among others). In particular, the results are brittle, require large corpora and do not support word sense disambiguation (which has repercussions for polysemous loanwords such as *kiwi*). In the context of studying language change, Bowern (2019) argues that word embeddings obscure critical data, overlooking the variation that is the input to change. We use word embedding plots for a different purpose here, namely, to help us glean the dominant semantic domain within which a hashtag occurs (given that we already know of its polysemy, following qualitative analyses of the data).

We trained word embeddings on the MLT corpus and identified the closest words in the semantic space to each of our hybrid hashtags. It was important to train embeddings on the MLT corpus rather than the HH sub-corpus because word embeddings work best with a large amount of training data. We implemented the *Word2Vec* algorithm (Mikolov et al., 2013) using Python's *Gensim* library (Rehurek and

TABLE 4 | Word-classes of the various hybrid hashtags in the HH sub-corpus.

Word-class	Hashtag example	Example of tweet containing hashtag	Num hashtags
Adjective Phrase (ADJP)	#kiwiproud	See you tonight Sydney City! Look for the wasted guy doing the haka. #KiwiProud hahahaha. [ID 523052566855184384]	3
Adverb Phrase (ADVP)	#kiwias	Usual weekend of sports entertainment resumes in NZ on @skysportnz this wkend! #SuperRugby #NRL #NBL #ALeague #kiwias #kiwi #kiwiana #sport! [ID 441819484534210560]	2
Common Noun Phrase (CNP)	#thehaka	So I don't know anything about #Rugby but I do know #TheHaka; Kiwi yr7 teacher had us do it :D Manly rugby boys doing it's a better view tho [ID 658053257416318976]	43
Formulaic Phrase (FMLA)	#kiaora4that	@tttrips Yeah…nah,not enuff gas bro but #kiaora4that anyway. He whakaaro Rangatira tena. [ID 272442027508117505]	5
Full Clause (CLAUSE)	#kiwiscanfly	Good luck to the kiwi triathletes racing in the European junior cup at Eton Dorney tomorrow @ETUtriathlon @TriathlonNZ #kiwiscanfly! #NZ [ID 373565680093630465]	6
Proper Noun Phrase (PNP)	#NZMaori	Going off to see the #nzmaori game today. Probability be more expat kiwis at the game than locals. [ID 396997290541326336]	6
Verb Phrase (VP)	#maorifyNZ	In order to #Maorifynz I will be swapping out my own Pakeha DNA with some spare Māori genes that Miriama Kamo has. [ID 90561843961147392]	13
N/A			3
Total			81

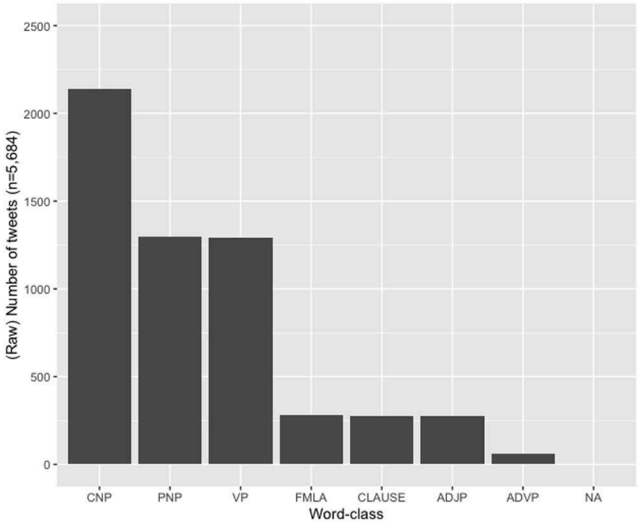


FIGURE 2 | Distribution across various word-classes in the hybrid hashtag set (CNP, common NP; PNP, proper NP; VP, verb phrase; FMLA, formulaic phrase; CLAUSE, full clause; ADJP, adjective phrase; ADVP, adverb phrase; NA, unsure).

TABLE 5 | Semantic domain of the various hybrid hashtags in the HH sub-corpus.

Semantic domain	Hashtag example	Example of tweet containing hashtag	Num hashtags
Flora and Fauna	#kiwiberries	I just discovered #kiwiberries, they are exactly what they sound like a small bite sized kiwi with no fuzz, best things ever! [ID 121230747351781377]	7
Generic	#kaitime	Honestly, no one can tell I'm Maori until they see me when there's seafood up for grabs… until then I'm pretty much plastic #kaitime [ID 91506535969021952]	2
Humor	#replacemovie quoteswithkiwi	my kiwi brings all the boys to the yard… #replacesongwordswithkiwi [ID 106461006527602689]	6
Māri culture	#keepinitreo	next week all orders at the drive thru in te reo maori #keepinitreo [ID 226445367913365504]	17
NZ Identity	#kiwislang	Caught myself saying something with a slight English accent today…I need to hear some kiwis ASAP #kiwisinlondon #kiwislang [ID 552521136127639554]	28
Sport	#kiwigold	@andreahewittnz does it again with a convincing first place at #ITU #GoldCoast #GoldCoastTri #kiwi #kiwigold [ID 850741519753596928]	20
N/A			1
Total			81

Sojka, 2010). After fine-tuning hyper-parameters, a CBOW architecture with negative sampling was chosen ($n = 5$), together with a window size of 15 and dimensionality of 200. This window size was chosen by maximizing the Mean Reciprocal Rank (MRR) of a list of chosen word-pairs (48 near-synonymous Māori/English word-pairs). The embeddings were then projected into two-dimensional space, using t-SNE (t-Distributed Stochastic Neighbor Embedding), a machine learning algorithm that preserves the distance between vectors when their dimensionality is reduced (see Maaten and Hinton, 2008).

In the resulting plots, the blue dot represents the target hybrid hashtag and the red dots represent the 40 closest words in the semantic space (those with the highest cosine similarity), which may consist of (native) English and/or Māori words. **Figures 4**, **5** show how these plots can help to identify the hashtag's semantic domain.

It is clear from **Figure 4** that the hashtag #proudkiwi pertains to sport. The semantic neighborhood includes the names of several famous New Zealand athletes (e.g., Mahe Drysdale, Andreea Hewitt, Lisa Carrington, George Bennett), specific sporting competitions (e.g., #London2012 Olympics), different sports in which New Zealanders excel (e.g., cycling, sailing, golf, rowing), references to "NZparalympics" and related hashtags (e.g., #EarnTheFern, #Gold).

Figure 5 relates to the hashtag #letssharegoodtereostories, and shows a number of Māori cultural terms, such as *#tereo* (the (Māori) language), *tupuna* (ancestors), *kaiako* (teacher) and *whaikōrero* (formal speech). Other words in the neighborhood are related to learning and promoting the Māori language (e.g., "immersion," "fluency," "bilingual_unit," "reconnect," "meaningful_dimensions," and "night_classes"), and/or to people's attitudes (e.g., "proud," *tu meke*/"too much"). From inspecting the plot, we can glean that this hashtag relates to the "Māori culture" semantic domain.

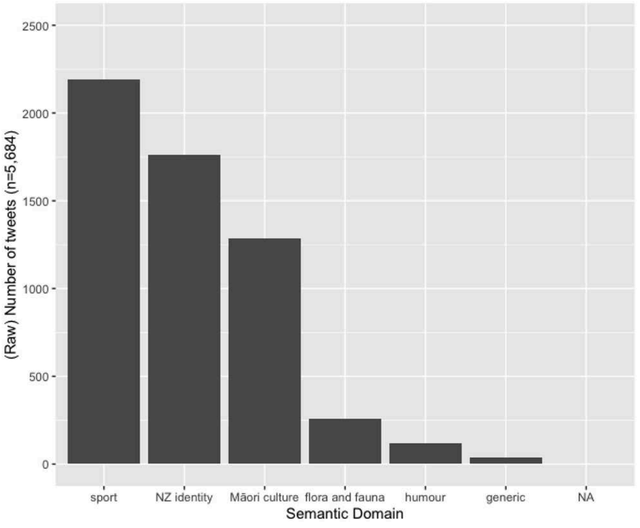


FIGURE 3 | Distribution across various semantic domains in the hybrid hashtag set.

4.2. Measuring Hashtag Survival/Life-Span

Given that the HH sub-corpus spans a period of 10 years, it is possible to investigate diachronic trends in the use of the hybrid hashtags extracted. Some of the hashtags rise more rapidly (e.g., #growingupkiwi, #youknowyoureakiwiwhen) or less rapidly (e.g., #kiwipride, #MāoriLanguageWeek), reach a peak and then decrease into disuse. Other hashtags have a cyclic life-span, whereby they are only used in specific months of the year recurrently, and not in other months (e.g., #TreatyofWaitangi). In general, as also noted by Maity et al. (2016), hashtags are highly transient and their life-span tends to be short. The hybrid hashtags in the HH sub-corpus are no exception to this trend.

We calculated Kendall Tau tests to check the status of the 81 hybrid hashtags in our set (by considering the more accurate counts of frequency per month), and found that 18 were statistically significantly increasing in use (#WaitangiDay, #proudkiwis, #letsshregoodtereostories, #kiwifruit, #hakarena, #kiwiproud, #kiwilove, #kiwias, #kiwisongs, #maorilanguage,

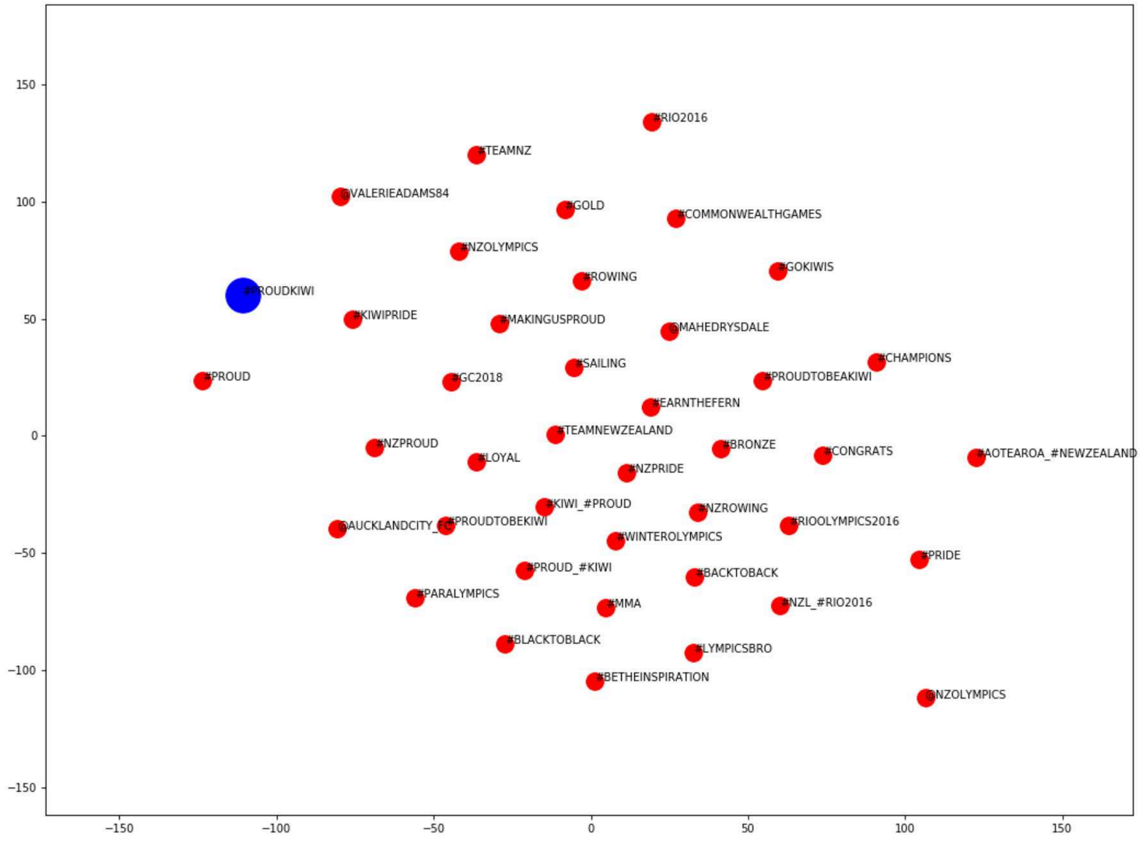


FIGURE 4 | Word embedding plot for the hashtag #proudkiwi.

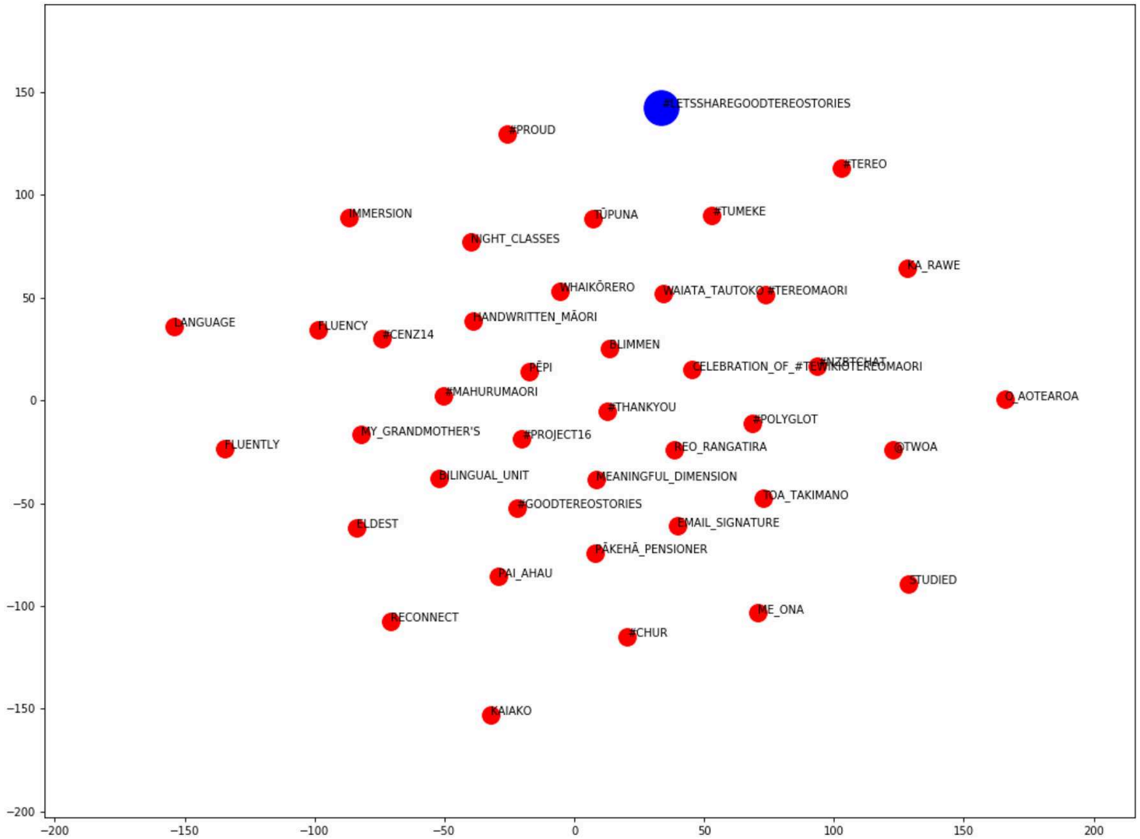


FIGURE 5 | Word embedding plot for the hashtag #letssharegoodtereostories.

#hakatime, #thehaka, #maoripride, #meanmaori, #kiaora4that, #proudmaori, #newkiwiburgersong, and #kiwiberries). The Kendall Tau test results for all 81 hashtags are reported in **Table S2**.

Studies which investigate hashtag survival use raw frequency of occurrence as a measure of the popularity of a given hashtag (e.g., Cunha et al., 2011; Tsur and Rappoport, 2012; Maity et al., 2016). There are few attempts to check these frequencies of use as they unfold over time—Maity et al. (2016) is a notable exception. In their work, Maity et al. (2016) track hashtag use by recording the (raw) number of occurrences of hashtags across weeks. However, one problem with this raw measure is that it does not distinguish between hashtags that occur across the same total number of weeks but which have a very different frequency distribution across those weeks. See, for example, the diachronic plots for the hybrid hashtags #huitweet and #kiaora4that in **Figure 6**[11].

Both these hashtags have a life-span of 5 (years), yet their use is very different within the 5-year period in which they occur. We propose an alternative measure of hashtag life-span (or survival) which takes into consideration both the duration that the hashtag is used for, as well as its relative activity or impact (i.e., how much it is used) in that period. Our notion of a hashtag's half-life is based on the idea of a word's half-life proposed by Pagel and Meade, which captures the point by which a given word-form has a 50% chance of being replaced by a non-cognate form (Pagel and Meade, 2006, 2018; Pagel et al., 2007). Analogously, the half-life of a hashtag captures the duration by which a given hashtag achieves 50% of its impact or activity (measured in frequency of use).

In practice, this measure can be operationalized separately for each hashtag, by calculating the amount of time it takes for a given hashtag to reach the half-point of the probability density function of its total observed frequency (during the period investigated). We did this in our data by using formulae in an Excel spreadsheet. The process is illustrated graphically in **Figure 7**, and mathematically, as follows. The hashtag in **Figure 7** has been simplified to show half-life in years (of which there are 10) for illustrative purposes—but we do not use years as our preferred time measure (we return to this further below). For now, let's consider the general process of calculating the half-life measure. The hashtag in **Figure 7** has a total frequency of use of 592 (occurrences), so it reaches its half-life at $592/2 = 296$ uses. The half-life measure is a temporal stamp, so we need to calculate the time it takes (starting from its very first use in the corpus in 2010) for the hashtag to reach the frequency of 296 occurrences (in 2014), which turns out to be 4 years (because $7_{2010} + 17_{2011} + 74_{2012} + 125_{2013} + 109_{2014} > 296$).

[11] We use number of years here rather than number of weeks or months for illustrative purposes, but the same argument holds for these measures.

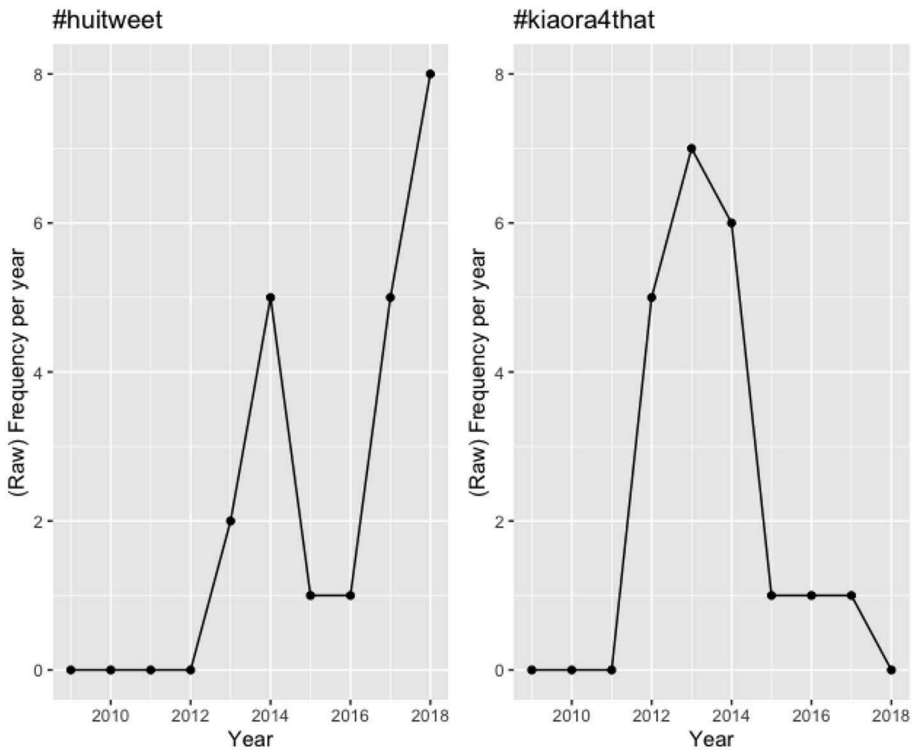


FIGURE 6 | Diachronic trend for #huitweet and #kiaora4that in the HH sub-corpus.

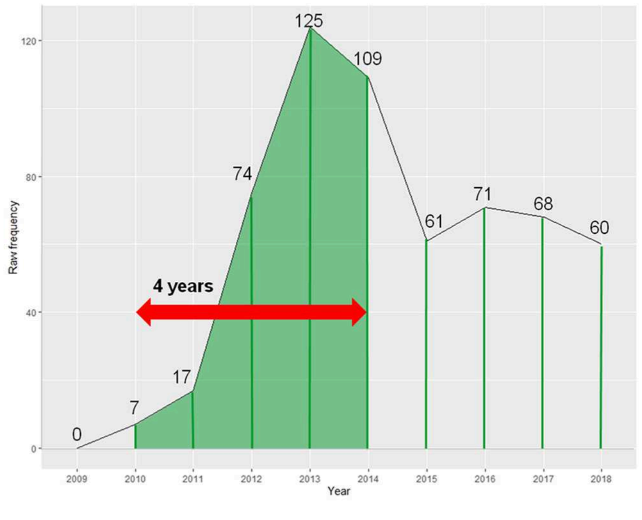


FIGURE 7 | Calculating the half-life of a hashtag.

Returning to **Figure 6**, #huitweet has a half-life of 4 years, whereas #kiaora4that has a half-life of 1 year, reflecting the different nature of their frequency distributions. We chose to measure half-lives of hybrid hashtags in our corpus across number of months in a bid to obtain the most fine-grained measurement (more accurate than years) while still avoiding data sparsity issues (which arose when considering number of weeks).

It is important to note that both existing measures of hashtag survival and the new measure we propose here (hashtag half-life) suffer from the drawback that they do not accurately capture the life-cycle of recently-coined hashtags. Current measures cannot say anything meaningful about the survival of such hashtags, given that we may not have seen their peak, or have been able to learn anything about the course of their use in the little time that they have existed on Twitter.

In our dataset, the half-life (estimated in number of months) values range between 0 months (for 13 distinct hashtags) and 79 months (for #kiwisdofly). See **Figure S2** for a frequency distribution of the various half-lives calculated for each of our 81 hybrid hashtags.

One obvious question to ask is whether there is any relationship between the various linguistic characteristics of the hashtags analyzed in the HH sub-corpus and their respective half-lives. **Figure 8** provides box-plot summaries of the various half-lives across each of these characteristics (semantic domain, word class, length of hashtag, and multiple variants).

The plots indicate that there are differences between the various types of hashtags (with respect to length, word-class, semantic domain, and whether or not hashtags are expressed by unique forms) and their respective half-lives. Since it is possible that all of these factors may influence a given hashtag's half-life (and, most likely, many other factors not coded here do too), we first used a Random Forest analysis implemented by the Boruta package in R (Kursa and Rudnicki, 2010) to check which factors are significantly associated with half-life scores. Boruta

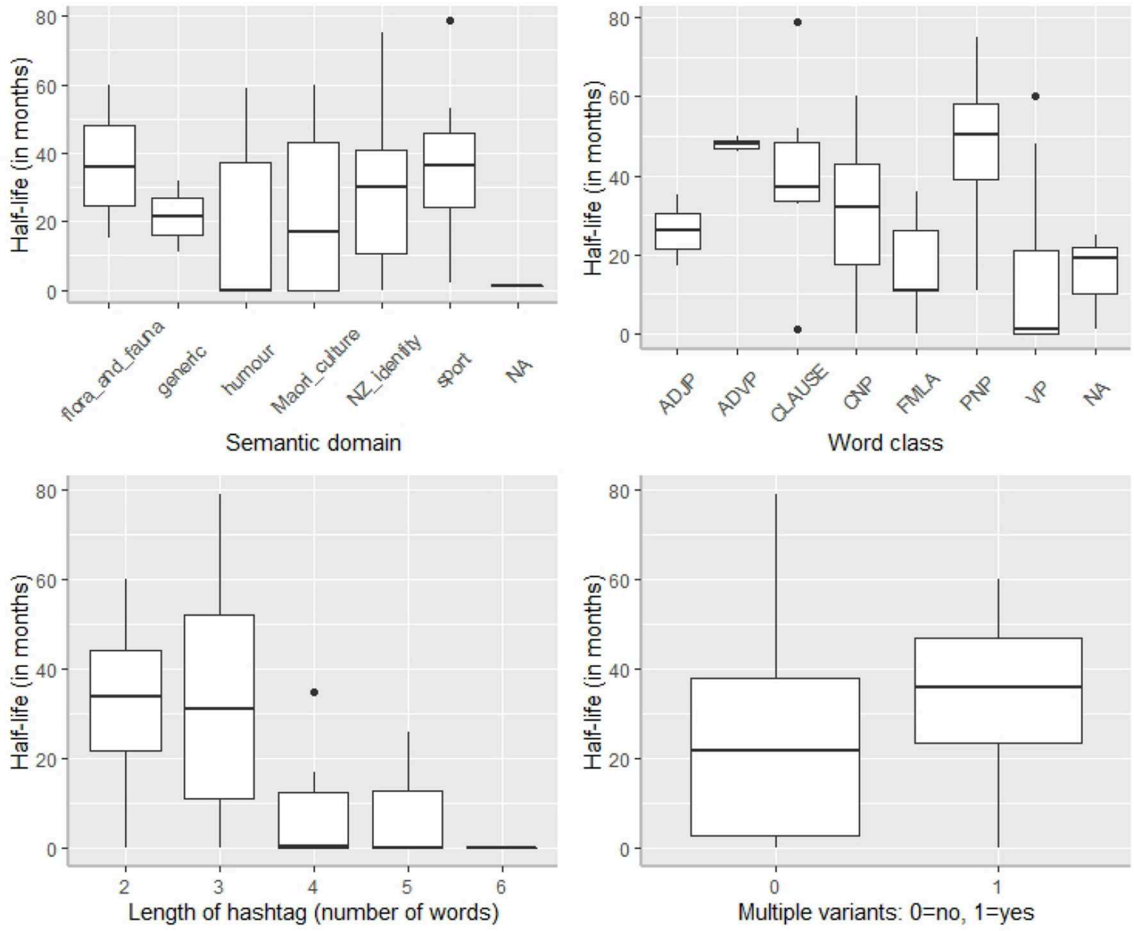


FIGURE 8 | Frequency distribution of half-lives of our 81 hybrid hashtags.

is a Random Forest technique which samples with replacement (unlike Conditional Inference Trees, see Baayen, 2008; Levshina, 2015).

Before running the Boruta function, we collapsed our word-class variable into two categories, namely, *nominal* (common and proper noun phrases) and *non-nominal* (all other classes: verb phrases, adverb phrases, adjective phrases, clauses, and formulaic hashtags). We also collapsed the semantic domain variable into four categories, namely, NZ identity, Māori culture, sport, and *other* (which includes humor, flora and fauna, and generic). This updated categorization system was adopted in order to ameliorate the under-representation problems of the original categories (for example, there were only two adjective-phrase hashtags). In addition to our four linguistic characteristics, we also included the hashtag, the user and the user frequency for each hashtag. This is because the same user is sometimes associated with multiple (distinct) hashtags, and different users will tweet the various hashtags with different frequencies. **Figure 9** gives the resulting plot. A description of each of these variables is given in **Table S3**.

We then built a step-up Generalized Mixed-Effects Model with a Quasi-Poisson distribution[12], modeling the half-life values obtained using the predictors that were deemed significant in the Boruta analysis (all except "user"). We thus included hashtag as a random variable, and the following remaining variables as fixed effects: semantic domain, length of hashtag, word class of hashtag, whether or not the hashtag had a unique form or multiple variants, and user frequency. The final minimal adequate model contained three factors: semantic domain, length of hashtag, and word class of hashtag, and a three-way interaction between these (see **Figure S4**, for further details). We inspected Cook's Distances and did not find outliers (see **Figure S4**). **Table 6** provides a detailed summary of the model. In general, increased hashtag length and non-nominal word-class are both associated with lower half-life scores; however, this effect is mediated by semantic domain of the hashtag. Non-nominal

[12] We first tried building a GLMM model with a Poisson distribution but this did not fit our data well (the overdispersion factor was 0.002004332), so we changed to a Quasi-Poisson distribution which performed much better (the overdispersion factor for the final minimal adequate model was 1.225681).

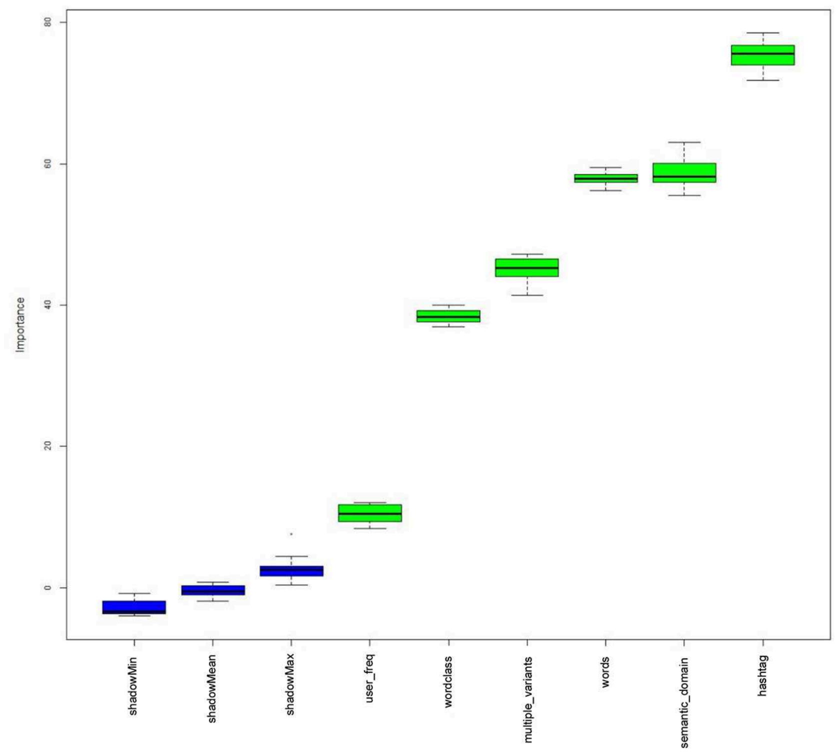


FIGURE 9 | Boruta plot showing the factors which are deemed to be relevant to half-life scores.

hashtags denoting sport or other concepts tend to have shorter half-lives compared to non-nominal hashtags denoting NZ identity. Conversely, nominal hashtags show the opposite trend: those denoting NZ identity have longer half-lives compared to those denoting sport or other concepts. Three-way interactions are notoriously difficult to interpret and these findings are only preliminary; more data are needed to confirm the trends.

It is important to emphasize that the models were not built for testing predictive power, but to test the influence of the variables. Given a particular hashtag, we would not expect the model to accurately predict its half-life; rather, the hypothesis tested here is whether or not a certain linguistic characteristic is statistically more likely to be associated with a higher half-life. Furthermore, due to practical constraints, the model lacks sociolinguistic predictors related to the users (such as gender, ethnicity, and status), which are also likely to influence hashtag life.

5. DISCUSSION

The previous section details our findings in relation to the set of hybrid hashtags found in the MLT corpus over the 10-year period investigated. While we cannot make any claims regarding the exhaustiveness of the Māori-English hybrid hashtags used on Twitter in general—our set of hybrid hashtags pertains only to the tweets obtained by means of the set of query words used to search the Twitter API—we believe that the data analyzed here can inform wider discussions of hashtags (beyond hybrid hashtags themselves) and current understanding of loanwords (as a linguistic and social phenomenon). We focus the discussion on three main issues.

5.1. Word-Formation in Hybrid Hashtags

As mentioned in section 2, there is divided opinion in the literature regarding the morphological word-formation process which gives rise to hashtags (see especially Caleffi, 2015; Maity et al., 2016). The most intuitive way to classify the formation of hashtags is by recourse to compounding, which is a problematic process in itself (see discussion in Bauer, 2017), but which appears to be among the most productive mechanism for creating new words in English. Certainly, some examples of hashtags in our data fit the compounding strategy well; see (12) and (13).

> (12) I love a good Kiwi accent. test = tist six = sex **#kiwiaccent** [ID 58156310386065408]

> (13) I remember going to the Zoo growing up and rarely seeing the Kiwis. Awesome news for the species! **#kiwibird** #kiwisandiegozoo… [ID 526886414118842369]

In (12), the common noun *Kiwi accent* parallels an existing productive compounding schema, e.g., British accent, Australian accent, American accent, as does the noun *kiwi bird* in (13), e.g., blackbird, bluebird, bellbird, tropicbird, secretarybird. These compounds are right-headed, as is typical of English compounds, and comprise a noun-noun combination, also a

TABLE 6 | Detailed summary of the GLMM model.

Predictor	Value	SE	DF	*t*-value	*p*-value
(Intercept)	2.811575	0.288647	4096	9.740528	0
words	0.066247	0.073783	62	0.897869	0.3727
wordclass_nonnominal	−0.2031	0.896049	62	−0.22666	0.8214
semantic_domain_New_Zealand_identity	**−9.48639**	**5.155466**	**62**	**−1.84006**	**0.0705**
semantic_domain_other	**23.6385**	**4.089416**	**62**	**5.780409**	**0**
semantic_domain_sport	0.130252	0.333948	62	0.390037	0.6978
words: wordclass_nonnominal	−1.08378	0.616854	62	−1.75695	0.0839
words: 3 semantic_domain_New_Zealand_identity	0.702501	2.574691	62	0.272849	0.7859
words: semantic_domain_other	**−12.5044**	**2.037805**	**62**	**−6.13622**	**0**
words: semantic_domain_sport	**−0.3489**	**0.11128**	**62**	**−3.1353**	**0.0026**
wordclass_nonnominal: semantic_domain_New_Zealand_identity	**15.62723**	**5.242248**	**62**	**2.981016**	**0.0041**
wordclass_nonnominal: semantic_domain_other	**−29.0146**	**4.253724**	**62**	**−6.82098**	**0**
wordclass_nonnominal: semantic_domain_sport	**2.066326**	**0.915875**	**62**	**2.256121**	**0.0276**
words: wordclass_nonnominal: semantic_domain_New_Zealand_identity	−3.87852	2.658266	62	−1.45904	0.1496
words: wordclass_nonnominal: semantic_domain_other	**13.65776**	**2.19546**	**62**	**6.220909**	**0**
words: wordclass_nonnominal: semantic_domain_sport	**3.046891**	**0.623952**	**62**	**4.883211**	**0**

Significant predictors are emphasized in bold.

highly utilized combination in English. The feature which makes these compounds distinctive is the combination of lexemes from distinct languages, Māori and English—but this type of combination has been documented as a productive word-formation strategy in other genres of New Zealand English (see Degani and Onysko, 2010).

However, compounding cannot account for hybrid hashtags that function as phrasal units exhibiting a productive syntactic frame, as evidenced by the variations we see in the hashtags' form [sometimes including the determiner, as in (15) and sometimes without it, as in (14)], but also by the existence of close alternative hashtags, such as (16) and (17). The lack of internal consistency violates one of the criteria proposed by Haspelmath (2011, p. 7) for word-hood. A second principle which appears to be potentially violated is that of potential pauses. Words are typically not able to include pauses (Haspelmath, 2011, p. 6). Of course, this is difficult to check in Twitter—a written language medium—but hashtags like #kiwiasbro, when uttered aloud are understood as separate words by speakers (*kiwi*, *as*, *bro*). This leads us to question the status of hashtags as words in the first place.

(14) So happy of our wee country! Best Olympics & now another gold, well done nz! So proud to be a kiwi #2012Olympics **#proudtobekiwi** #nzolympics [ID 234994140339900416]

(15) Double Gold! No voice and one bloody proud kiwi! #GoKiwi @nzolympics **#proudtobeakiwi** [ID 231354255653621760]

(16) #kiakaha today @RealStevenAdams in your first #NBA start. Play hard, enjoy the game. **#kiwiproud** [ID 400777324062187521]

(17) **#ProudKiwi** im a proud kiwi rt if you are to favorite if you from auckland [ID 235017500000133121]

Even more problematic hashtags are those which span entire clauses, as in (18) and (19). The complex internal structure of clausal hashtags is also noted by Caleffi (2015) and forms the main evidence for her proposal that hashtags represent a completely distinct word-formation process, which she terms *hashtagging*.

(18) **#kiwisareawesomepeople** for protecting their native animals like kiwis,kea,kekapo,weka,morepork [ID 25866163769]

(19) Its kinda depressing that I might be allergic to Kiwi. **#ilovekiwi** [ID 474333666814877696]

The meanings of hashtags in the examples above can only be decoded by taking into consideration the meaning and syntactic role of the individual words comprising the hashtag, in the same manner as any other clause in English. The only difference is the orthographic appearance of the hashtag, which uses the "#" symbol and lacks spaces between words. Moreover, the syntactic structure of the hashtag can be expanded to richer and more elaborate hashtags, e.g., #ilovefunnykiwis or #heloveskiwis, to create novel hashtags, in a highly productive fashion, reminiscent of typical English phrasal structures.

We question the status of hashtags as words and suggest that hashtags are, at best, artificial words, and therefore outside the scope of the usual morphological formation processes that would typically underpin the formation of (legitimately) new words in a language system[13]. Given their function in discourse, these units must "look," orthographically, like individual words (by having spaces removed between their components) in order to facilitate searchability and discovery by other online community members. However, linguistically, we argue that they should not be analyzed as actual words because they are derived from a number of distinct processes (some of which are indeed akin to compounding, while others are not), and interpreted by recourse to analysis of the individual components within each use.

5.2. Function of Hybrid Hashtags in Discourse

Previous work on loanwords identifies a number of linguistic and non-linguistic reasons for the adoption of lexical material from one language into another. These include filling lexical gaps in the receiver language or lexical gaps of bilingual speakers, economy of expression, expression of identity, language regard, and so on (Poplack, 2018, chapter 11 and others).

One factor which has been relatively under-represented in the literature on loanwords (but see Macalister, 2002 for a handful of examples from New Zealand English) is the dimension of humor and language play. Language play and creative uses of linguistic resources (see Zirker and Winter-Froemel, 2015 and papers cited within) have been documented in monolingual contexts of word formation (Renner, 2015) and in English-German bilingual puns (Stefanowitsch, 2002; Knospe, 2015), but to our knowledge, they are largely absent from studies of loanwords. Given the link between creativity and bilingualism (see overview in Kharkhurin, 2015), it is perhaps not surprising that loanwords illustrate creative language use and language play.

We found that Twitter is a particularly rich genre for investigating language play in loanword use. Although we devised a specific semantic function category to include hybrid hashtags whose primary function is that of invoking humor, many of the other uses of hybrid hashtags appeared to also exhibit language play and humorous undertones, even if this was not their primary function. As an illustration of this phenomenon, consider example (20).

(20) it's time to start focusing on regional economic development for our whanau and runanga says @ngaitahu **#honeyhui** [ID 760990045389987840]

In (20), the Māori word *hui* is roughly translated in English as "meeting" or "gathering." The hybrid hashtag #honeyhui is used in the above tweet by MBIE (*The Ministry of Business, Innovation and Employment*) as a creative reference to the English concept of a "working bee," bringing a light-hearted touch to an otherwise serious and controversial effort to improve the economic situation of regional councils and (New Zealand) families. The councils and families in question are referenced by means of Māori loanwords (the word *whānau* refers to family and extended family members, and the word *runanga* refers to a council). The use of Māori loanwords for these concepts is socially meaningful because it invokes an inclusive practice, emphasizing the fact that the effort aims to improve the economic development of all regional councils and families; the use of Māori loanwords references those councils and families predominantly made up of Māori (and thereby explicitly referencing groups which might have previously been marginalized from such an effort). The discourse function of the hybrid hashtag #honeyhui has less to do with categorizing the tweet or with signaling group affiliation, and more to do with bringing together two distinct worldviews and points of reference, in a suggested unified action to improve economic development. The hashtag functions as a softening device (achieved through light-hearted humor), aimed at defusing tension in a delicate and socially-charged situation. Other phenomena unique to computer-mediated communication, such as emojis, can play a similar role in the diffusion of tension (for further discussion, see Evans, 2017). The example shows the richness of meaning that can be derived from loanword use and the different layers of interpretation arising from this use.

Additional examples of hashtags with humorous undertones can be seen in the use of the hashtags #youknowyoure(a)kiwiwhen and #growingupkiwi, in examples (21) and (22), respectively. Both these tags primarily discuss issues of New Zealand identity (and are categorized as such in our analysis), but they also bring in a playful dimension. In (21), the user laments the Marmite shortage that occurred when Sanitarium ceased production of Marmite, due to factory damage caused by the 2011 Christchurch earthquake. This shortage caused an uproar in the New Zealand community because the New Zealand brand of Marmite is seen an icon of kiwi culture. The hashtag #youknowyoure(a)kiwiwhen facilitates the user's attempt to poke fun at the problem of grieving the loss of marmite by implying that only a New Zealander would understand this loss and by hinting (implicitly) that the magnitude or validity of this loss is underestimated by those who are not New Zealanders.

[13] We are grateful to Laurie Bauer for his input which shaped this proposal.

(21) **#youknowyourekiwiwhen** you grieve the loss of marmite [ID 427393399855923200]

(22) **#growingupkiwi** being a skinny white kid in a Primary school Kapa Haka group [ID 621264554266243072]

In (22), #growingupkiwi is similarly used to focus attention on the experience of being a New Zealander, and presents this experience as distinct and perhaps misunderstood by outsiders. *Kapa haka* groups are traditional Māori performance groups, typically made up of Māori children, but in recent years, children of European descent have started to join in too (referenced by the comment about being the "skinny white kid" among the predominantly dark-skinned Māori children in the group).

Unlike #honeyhui, the hashtags #youknowyoure(a)kiwiwhen and #growingupkiwi are humorous not because of word-play, but because they describe relatable, shared experiences of being a New Zealander and being raised in New Zealand.

The examination of Twitter data may be more conducive to discovering creative uses of loanwords compared to other genres because of the informal and potentially anonymous[14] nature of the posts. Compared to newspaper language which involves ample editing and scrutiny, or even recorded conversational data, in which speakers are aware of the fact that they are being recorded, Twitter affords a rapid and uncensored window into off-the-cuff language use.

A second observation to be made about the function of hashtags on Twitter is that, as argued in section 4.1, while it is true that hashtags can and do function as affiliative tags and categorizing and community-building devices at a macro-level (see discussion of the hashtag #banthehaka as a discoverable tag for joining the debate about the performance of the haka in rugby matches), they also have a purely semantic dimension, expressing actual linguistic content, at a micro-level. We hope to have shown that, while the two roles can sometimes fruitfully co-exist, there are also cases where one role is foregrounded to the partial or complete exclusion of the other. For instance, the semantic content of #honeyhui is more important than the categorizing function in example (20), and the affiliative role is primary for #banthehaka in example (8), rendering the semantic content of the hashtag obsolete.

5.3. Integratedness of Loanwords in Receiver Language

One final observation we make relates to what Twitter and hybrid hashtags might be able to tell us about loanword integration. The question of how to determine the entrenchment of loanwords within a receiver language is a longstanding problem (see discussion in Turpin, 1998; Jones, 2005; Zenner et al., 2014; Levendis and Calude, 2019). This issue is particularly problematic in the context of English as a receiver language because typical ways of establishing entrenchment of loanwords involve examining morphological and phonological integration of loanwords in the adoptive language, and English has a distinct lack of morphological marking[15]. Additionally, some studies cite listedness as a factor in establishing entrenchment (Stammers and Deuchar, 2012, p. 631), but recent work casts some doubt as to whether that is a robust measure for Māori loanwords in (New Zealand) English (Levendis and Calude, 2019).

Given the time and effort costs involved in obtaining the spoken language data required to tap into phonological integration, morphological integration remains a key factor in determining loanword entrenchment. As regards English, one of the few morphological strategies for signaling entrenchment of a loanword cited in the literature is plural marking (on nouns). However, for prescriptive reasons, this strategy has been actively discouraged in New Zealand with regard to Māori loanwords (see Davies and MacLagan, 2006, p. 90). Interestingly, there is one loanword which appears to be exempt from this "rule," namely the loanword *kiwi* (*kiwis* does not appear to attract criticism)—this exemption is likely a sign of entrenchment in itself because it points to the fact that many speakers of New Zealand English are no longer conscious of the fact that *kiwi* is borrowed from Māori.

Our corpus of hybrid hashtags shows two further possible sources of evidence for loanword entrenchment, namely the use of loanwords in hybrid hashtags and the use of derivation. Because hybrid hashtags involve loanwords that have been found to be very frequent in other corpora (see discussion in section 4.1), it seems reasonable to assume that the presence of a hybrid hashtag involving a given loanword can be taken to be a sign of entrenchment of that loanword in English. Secondly, our corpus exhibits some (albeit few) examples of loanwords used with productive English derivational suffixes, see examples (23) and (24).

(23) I'm outnumbered in this café by French speakers. Rather cool. But it'd be better to only hear Te Reo. **#maorifynz** [ID 98119407166955520]

(24) Using te reo tongue-twisters makes even the simplest acting warm-up games tricky (and hilarious). **#maorifynz** [ID 169695510075158530]

Both the presence of derivation and the use of loanwords in hybrid hashtags are predictors of entrenchment; however, the absence of these is not necessarily an indicator of a lack of entrenchment.

6. CONCLUSION

This paper reports findings related to a set of productive hybrid hashtags, made up of lexical components from two separate languages, namely, a minority, indigenous language (te reo Māori) and a dominant lingua franca (English). The hybrid hashtags are extracted from a diachronic corpus of tweets, over a 10-year period between 2009 and 2018, and analyzed using

[14] Some people do not use their real names on Twitter.

[15] There is a wealth of work being done on phonological integration of loanwords, too large to cite here, but for a recent and meticulous study of phonological integration of Māori loanwords in New Zealand English, see Hashimoto, 2019 and references cited within.

a combination of descriptive and quantitative tools. The main contributions of this paper are as follows:

- described semantic and syntactic categories of hybrid hashtags, as well as their functions in discourse;
- proposed and operationalized a new metric for measuring the life-cycle of a hashtag, a hashtag's half-life;
- proposed additional criteria for measuring loanword morphological integration;
- studied the role of loanwords from te reo Māori in (primarily, New Zealand) English and society.

We find that Twitter constitutes a rich source of investigating loanwords and language-mixing phenomena, as well as informal, creative language use. The data analyzed show that hybrid hashtags are extremely versatile with regard to their length, semantic function and word-class, encompassing various types of each. Given that hybrid hashtags appear to be composed of loanwords which are known to be highly productive in other genres, we argue that the presence of a loanword in a hybrid hashtag could be a reliable predictor of loanword entrenchment.

Concerning hashtags more generally, the internal versatility of the hashtags we analyzed and the need for decomposition in order to decode their semantic content point to the fact that hashtags are best regarded as artificial words (and not true words), which cannot be derived through compounding or other traditional word-formation processes. Secondly, their function in discourse is of a dual nature: on the one hand, they have a micro-discourse role in which they carry semantic meaning (this can be downgraded or altogether canceled if it conflicts with their wider discourse function), and at the same time, they have a macro-discourse role in which they act as community-forming or categorizing devices (this can similarly be downgraded in favor of their micro-discourse role).

One cited benefit of analysing language on Twitter is the rapid nature of change, observable within a shorter time frame than linguists are typically used to Grieve et al. (2018), and hashtags, in particular, constitute a perfect example of a fast-changing, highly transient linguistic phenomenon. We problematize current measures of hashtag life-span, which take into consideration duration of existence, but neglect to measure overall impact, and propose a new measure of hashtag life-span, namely, the hashtag's *half-life*. We build statistical models which show that there are associations between linguistic properties of the hashtags analyzed and their half-lives, although these models currently suffer from several limitations (they are missing factors related to the content of the tweets containing the hashtags and features related to the user, such as gender and ethnicity)—limitations which we leave for future work.

DATA AVAILABILITY STATEMENT

The dataset generated for this study can be found on the Kiwi Words website at waikato.github.io/kiwiwords/hh_corpus.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This research was funded by the Royal Society Marsden Fund (NZ).

ACKNOWLEDGMENTS

The authors would like to thank the editors, as well as the reviewers of this Special Issue. We thank Mark Pagel for fruitful discussions of half-life, and Martin Schweinberger for his generous help with the statistical modeling. Any remaining errors are our own.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00015/full#supplementary-material

REFERENCES

Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., and Odusami, M. (2019). A review of soft techniques for sms spam classification: methods, approaches and applications. *Eng. Appl. Artif. Intell.* 86, 197–212. doi: 10.1016/j.engappai.2019.08.024

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Backus, A. (2013). "A usage-based approach to borrowability," in *New Perspectives on Lexical Borrowing*, eds E. Zenner and G. Kristiansen (Boston, MA: Mouton de Gruyter Berlin), 19–39.

Bauer, L. (2017). *Compounds and Compounding, Vol. 155*. Cambridge: Cambridge University Press.

Bowern, C. (2019). "Semantic change and semantic stability: variation is key," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (Florence: Association for Computational Linguistics), 48–55.

Caleffi, P.-M. (2015). The 'hashtag': a new word or a new rule? *SKASE J. Theor. Linguist.* 12, 46–69.

Calude, A., Harper, S., Miller, S., and Whaanga, H. (2019a). Detecting language change: māori loanwords in a diachronic topic-constrained corpus of New Zealand English newspapers. *Asia Pac. Lang. Variat.* 5, 109–138. doi: 10.1075/aplv.00003.cal

Calude, A., Stevenson, L., Whaanga, H., and Keegan, T. T. (2019b). The use of māori words in National Science Challenge online discourse. *J. R. Soc. N. Z.* 1–18. doi: 10.1080/03036758.2019.1662818

Calude, A. S., Miller, S., and Pagel, M. (2017). Modelling loanword success–a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguist. Linguist. Theory*. 15, 1–38. doi: 10.1515/cllt-2017-0010

Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2011). "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," in *Proceedings of the Workshop on Languages in Social Media*, (Portland, OR: Association for Computational Linguistics), 58–65.

Daly, N. (2007). Kūkupa, koro, and kai: the use of Māori vocabulary items in New Zealand English children's picture books. *New Zealand Eng. J.* 21, 20–33.

Daly, N. (2016). Dual language picturebooks in english and māori. *Bookbird* 54, 10–17. doi: 10.1353/bkb.2016.0092

Davies, C., and MacLagan, M. (2006). Māori words–read all about it: testing the presence of 13 māori words in four New Zealand newspapers from 1997 to 2004. *Te Reo*. 49, 73–99.

de Bres, J. (2006). Maori lexical items in the mainstream television news in New Zealand. *N. Z. Engl. J.* 20:17.

Degani, M. (2010). The Pakeha myth of one New Zealand/Aotearoa: an exploration in the use of Maori loanwords in New Zealand English. *From Intl. Local English–and Back Again* 165–196.

Degani, M., and Onysko, A. (2010). Hybrid compounding in New Zealand English. *World Engl.* 29, 209–233. doi: 10.1111/j.1467-971X.2010.01639.x

Evans, V. (2015). *#Language: Evolution in the Digital Age Their Use of the Hashtag Shows That Under 13s Are at the Vanguard of Linguistic Innovation.* The Guardian.

Evans, V. (2017). *The Emoji Code: The Linguistics Behind Smiley Faces and Scaredy Cats*. New York, NY: Picador.

Field, F. W. (2002). *Linguistic Borrowing in Bilingual Contexts, Vol. 62*. Amsterdam: John Benjamins Publishing.

Grieve, J., Nini, A., and Guo, D. (2017). Analyzing lexical emergence in Modern American English online. *Engl. Lang. Linguist.* 21, 99–127. doi: 10.1017/S1360674316000113

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on American social media. *J. Engl. Linguist.* 46, 293–319. doi: 10.1177/0075424218793191

Hashimoto, D. (2019). *Loanword Phonology in New Zealand English: Exemplar Activation and Message Predictability*. Doctoral dissertation, Canterbury University, Christchurch, New Zealand. Retrieved from http://hdl.handle.net/10092/16634

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguist.* 45, 31–80. doi: 10.1515/flin.2011.002

Haspelmath, M., and Tadmor, U. (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Walter de Gruyter.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language* 26, 210–231. doi: 10.2307/410058

Hay, J. (2018). "What does it mean to "know a word?"," in *Language and Society Conference of New Zealand in November 2018 in Wellington, New Zealand* (Wellington).

Holmes, J., Johnson, G., and Vine, B. (1998). *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies; Victoria University of Wellington.

Jones, M. C. (2005). Some structural and social correlates of single word intrasentential code-switching in Jersey Norman French. *J. French Lang. Stud.* 15, 1–23. doi: 10.1017/S0959269505001894

Jurgens, D., Dimitrov, S., and Ruths, D. (2014). "Twitter users# codeswitch hashtags!# moltoimportante# wow," in *Proceedings of the First Workshop on Computational Approaches to Code Switching* (Doha), 51–61.

Keegan, T. T., Mato, P., and Ruru, S. (2015). Using Twitter in an indigenous language: an analysis of Te Reo Māori tweets. *AlterNative* 11, 59–75. doi: 10.1177/117718011501100105

Kennedy, G., and Yamazaki, S. (1999). The influence of Maori on the Nw Zealand English lexicon. *Lang. Comput.* 30, 33–44.

Kharkhurin, A. V. (2015). *Bilingualism and Creativity*. Chichester: Wiley Online Library.

Knospe, S. (2015). *A Cognitive Model for Bilingual Puns.* Berlin: De Gruyter.

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Lee, C., and Chau, D. (2018). Language as pride, love, and hate: archiving emotions through multilingual instagram hashtags. *Discourse Context Media* 22, 21–29. doi: 10.1016/j.dcm.2017.06.002

Levendis, K., and Calude, A. (2019). Perception and flagging of loanwords–a diachronic case-study of māori loanwords in new zealand english. *Ampersand* 6:100056. doi: 10.1016/j.amper.2019.100056

Levshina, N. (2015). *How To Do Linguistics With R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company.

Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Macalister, J. (2002). Maori loan words and New Zealand humour. *NZ Words* 6, 3–6.

Macalister, J. (2006). The Maori presence in the New Zealand English lexicon, 1850–2000: evidence from a corpus-based study. *Engl. World-Wide* 27, 1–24. doi: 10.1075/eww.27.1.02mac

Macalister, J. (2009). Investigating the changing use of Te Reo. *NZ Words* 13, 3–4.

Maity, S. K., Saraf, R., and Mukherjee, A. (2016). "#Bieber+#Blast=#Bieberblast: early prediction of popular hashtag compounds," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York, NY: ACM), 50–63.

Matras, Y. (2009). *Language Contact*. Cambridge: Cambridge University Press.

McCallum, A., and Nigam, K. (1998). "A comparison of event models for naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization, Vol. 752* (Stroudsburg, PA: Citeseer), 41–48.

McMonagle, S., Cunliffe, D., Jongbloed-Faber, L., and Jarvis, P. (2019). What can hashtags tell us about minority languages on twitter? a comparison of# cymraeg,# frysk, and# gaeilge. *J. Multiling. Multicult. Dev.* 40, 32–49. doi: 10.1080/01434632.2018.1465429

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems* (Stateline, NV), 3111–3119.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.

Onysko, A., and Calude, A. (2013). "Comparing the usage of Māori loans in spoken and written New Zealand English: a case study of Māori, Pākehā, and Kiwi," in *New Perspectives on Lexical Borrowing: Onomasiological, Methodological, and Phraseological Innovations* (De Gruyter), 143–170.

Page, R. (2012). The linguistics of self-branding and micro-celebrity in Twitter: the role of hashtags. *Discour. Commun.* 6, 181–201. doi: 10.1177/1750481312437441

Pagel, M., Atkinson, Q., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449, 717–720. doi: 10.1038/nature06176

Pagel, M., and Meade, A. (2006). *Estimating Rates of Lexical Replacement on Phylogenetic Trees of Languages*. Cambridge: McDonald Institute for Archaeological Research.

Pagel, M., and Meade, A. (2018). The deep history of the number words. *Philos. Trans. R. Soc. B Biol. Sci.* 373, 1–9. doi: 10.1098/rstb.2016.0517

Poplack, S. (2018). *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford: Oxford University Press.

Poplack, S., and Sankoff, D. (1984). Borrowing: the synchrony of integration. *Linguistics* 22, 99–136. doi: 10.1515/ling.1984.22.1.99

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26, 47–104. doi: 10.1515/ling.1988.26.1.47

Preston, D. R. (2013). The influence of regard on language variation and change. *J. Pragmat.* 52, 93–104. doi: 10.1016/j.pragma.2012.12.015

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rehurek, R., and Sojka, P. (2010). "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta: Citeseer).

Renner, V. (2015). *Lexical Blending as Wordplay*. Berlin: Mouton de Gruyter.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad: ACM), 695–704.

Sapir, E. (2004). *Language: An Introduction to the Study of Speech*. New York, NY: Courier Corporation.

Stammers, J. R., and Deuchar, M. (2012). Testing the nonce borrowing hypothesis: counter-evidence from English-origin verbs in Welsh. *Bilingualism* 15, 630–643. doi: 10.1017/S1366728911000381

Stefanowitsch, A. (2002). Nice to miet you: bilingual puns and the status of English in Germany. *Intercult. Commun. Stud.* 11, 67–84.

Trye, D., Calude, A., Bravo-Marquez, F., and Keegan, T. T. (2019). "Māori loanwords: a corpus of New Zealand English tweets," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (Florence: Association for Computational Linguistics), 136–142.

Tsur, O., and Rappoport, A. (2012). "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, WA: ACM), 643–652.

Turpin, D. (1998). 'Le français, c'est le last frontier': the status of english-origin nouns in Acadian French. *Int. J. Bilingual.* 2, 221–233. doi: 10.1177/136700699800200206

Weinrich, U. (1953). *Languages in Contact. Findings and Problems*. New York, NY: Mouton.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Zappavigna, M. (2011). Ambient affiliation: a linguistic perspective on twitter. *New Media Soc.* 13, 788–806. doi: 10.1177/1461444810385097

Zenner, E., Rosseel, L., and Calude, A. S. (2019). The social meaning potential of loanwords: empirical explorations of lexical borrowing as expression of (social) identity. *Ampersand* 6:100055. doi: 10.1016/j.amper.2019.100055

Zenner, E., Speelman, D., and Geeraerts, D. (2014). Core vocabulary, borrowability and entrenchment: a usage-based onomasiological approach. *Diachronica* 31, 74–105. doi: 10.1075/dia.31.1.03zen

Zirker, A., and Winter-Froemel, E. (2015). *Wordplay and Its Interfaces in Speaker-Hearer Interaction: An Introduction.* Berlin: Mouton de Gruyter.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

METHODS
published: 15 April 2020
doi: 10.3389/frai.2020.00010

Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study

Anne-Sophie Ghyselen[1]*, *Anne Breitbarth*[1], *Melissa Farasyn*[1], *Jacques Van Keymeulen*[1,2] *and Arjan van Hessen*[3]

[1] *Department of Linguistics, Ghent University, Ghent, Belgium,* [2] *Variaties VZW, Umbrella Organisation for Dialects and Oral Heritage, Brussels, Belgium,* [3] *Human Media Interaction, Faculty of Electrical Engineering, Mathematics & Computer Science, University of Twente, Enschede, Netherlands*

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Karlien Franco,
KU Leuven, Belgium
Remco Knooihuizen,
University of Groningen, Netherlands

***Correspondence:**
Anne-Sophie Ghyselen
annesophie.ghyselen@ugent.be

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *17 September 2019*
Accepted: *28 February 2020*
Published: *15 April 2020*

Citation:
Ghyselen A-S, Breitbarth A, Farasyn M, Van Keymeulen J and van Hessen A (2020) Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study. Front. Artif. Intell. 3:10. doi: 10.3389/frai.2020.00010

This paper discusses how the transcription hurdle in dialect corpus building can be cleared. While corpus analysis has strongly gained in popularity in linguistic research, dialect corpora are still relatively scarce. This scarcity can be attributed to several factors, one of which is the challenging nature of transcribing dialects, given a lack of both orthographic norms for many dialects and speech technological tools trained on dialect data. This paper addresses the questions (i) how dialects can be transcribed efficiently and (ii) whether speech technological tools can lighten the transcription work. These questions are tackled using the Southern Dutch dialects (SDDs) as case study, for which the usefulness of automatic speech recognition (ASR), respeaking, and forced alignment is considered. Tests with these tools indicate that dialects still constitute a major speech technological challenge. In the case of the SDDs, the decision was made to use speech technology only for the word-level segmentation of the audio files, as the transcription itself could not be sped up by ASR tools. The discussion does however indicate that the usefulness of ASR and other related tools for a dialect corpus project is strongly determined by the sound quality of the dialect recordings, the availability of statistical dialect-specific models, the degree of linguistic differentiation between the dialects and the standard language, and the goals the transcripts have to serve.

Keywords: dialect, transcription, corpus research, ASR, respeaking, forced alignment, dutch, Flanders

INTRODUCTION

In the history of dialectological research, corpus research has long been scarce. Dialect atlases and dictionaries traditionally build on survey data and/or introspective data (native speaker judgments), rather than on databases of spontaneous speech samples. The reasons for the popularity of these survey and introspective data are quite obvious: (1) on the basis of elicitation and introspection, the diverse aspects of a dialect's lexicon, phonology, morphology, and/or syntax can be studied more systematically, by restricting the focus to controlled conditions (cf. Cornips and Poletto, 2005), and (2) the collection and analysis of elicited/introspective data are also less time-consuming than dialect corpus building and analysis. The restriction to predefined conditions, however, while making research efficient, replicable, and comparable, is also a major limitation.

Dialect corpus research has clear advantages over elicited data here: analyzing spontaneous speech not only allows insight into the functional strength of dialect features in real life but also makes possible a more thorough study of dialect phenomena conditioned by discourse or register, phenomena that might remain unnoticed in survey data. Not in the least, it allows for the serendipitous discovery of phenomena that previously escaped attention and are therefore not considered in the construction of surveys.

In usage-based approaches (Kemmer and Barlow, 2000; Bybee, 2010) as much as in more formalist (especially historical) research (cf. contributions in Jonas et al., 2011; and Mathieu and Truswell, 2017), corpus analysis has strongly gained in popularity (cf. Szmrecsanyi and Anderwald, 2018), as frequency data are a way to uncover/reconstruct the linguistic knowledge underlying the usage, and to study contextual factors affecting it. This development is also fostered by the availability of Automated Speech Recognition (ASR) tools and Natural Language Processing (NLP) software facilitating automated audio and text annotation. Remarkably, however, *dialect* corpora are still relatively scarce, especially when the term 'dialect' is interpreted in the 'traditional' sense as regionally determined language varieties that differ at multiple structural levels—phonetic, phonological, morphological, lexical, syntactic, and/or semantic—from other dialects and the 'overarching' standard language (cf. Trudgill, 1999, p. 5; Boberg et al., 2018, p. 4–5).[1] A number of factors account for this scarcity. First, dialects are generally spoken in informal/private domains, making it challenging to collect samples of these language varieties. In contrast to standard language corpora, one cannot partly rely on 'public' speech settings, such as news broadcasts, TV shows, or parliament debates for data collection. Secondly, as ASR and NLP tools are usually trained on standard language data, it can be quite challenging to apply these tools to dialect data. As such, transcribing or annotating dialect data usually requires more manual work than standard language data (or regionally accented language use). Even when disregarding the functioning of ASR and NLP tools, the process of putting speech to text—the first essential step in the building of speech corpora—is much more challenging for dialect recordings than for standard languages, as for many dialects orthographic norms are not available.

Interestingly, the transcription problem in dialect corpus research has received little scientific attention, which is strange given the increased interest in transcript-based research the last decades. In this paper, we aim at filling this methodological gap by addressing the questions (i) how dialects can be transcribed efficiently and (ii) whether NLP tools can lighten the transcription work. These questions will be tackled using the Southern Dutch dialects (SDDs) as case study, i.e., the dialects spoken in (i) Dutch-speaking Belgium, (ii) the three southern provinces of the Netherlands (Limburg, Noord-Brabant, and Zeeland), and (iii) the Flemish-speaking dialect region in France.[2] The discussion is based on the results of a pilot project laying the foundations for a large-scale Corpus of SDDs. The pilot project, which focused on the dialect collection *Stemmen uit het Verleden* ('Voices from the past,' Ghent University)[3], aimed at developing a transcription protocol and an annotation pipeline and establishing benchmarks for the transcription, correction, and annotation of Dutch dialect recordings.

TOWARD A CORPUS OF SDDs

The SDDs have been shown to have a number of striking typological characteristics (see, e.g., De Vogelaer, 2008; De Schutter, 2009; Taeldeman and De Wulf, 2010; Swanenberg and van Hout, 2013; Breitbarth and Haegeman, 2014), with dialects diverging phonologically, morphologically, syntactically, and lexically from both the Dutch standard language[4] and each other. In the light of the so-called "delayed" standardization process in Flanders (Vandekerckhove, 2009, p. 75), dialect leveling processes have set in quite late (compared to other European speech communities), and hence, dialects still often vary from village to village or from city to city. This dialect diversity is interesting for language-historical research, as the SDDs form a missing link in the language history since Middle Dutch: the SDDs played only a minor role in the standardization processes mainly going out from the northern provinces since the seventeenth century, and were hardly affected by them (cf. Willemyns, 2003).

Much of the more recent research into the SDDs is either based on the big dialect atlases of Dutch, i.e., the *Fonologische Atlas van de Nederlandse Dialecten* (FAND, 1998/2000/2005; 'Phonological Atlas of the Dutch Dialects'), the *Morfologische Atlas van de Nederlandse Dialecten* (MAND, 2005/2009; 'Morphological Atlas of the Dutch Dialects'), and the *Syntactische Atlas van de Nederlandse Dialecten* (SAND, 2005/2008; 'Syntactic Atlas of the Dutch Dialects'), which are based on elicited data, or on introspective data (native speaker judgments). As already discussed in the *Introduction*, there are a number of problems with these methods when it comes to linguistic research, especially into the syntax of Flemish dialects. For example: contrary to Standard Dutch, some SDDs can have the verb as the third constituent in the clause [cf. (1)] instead of the second one, if the clause is introduced by an adverbial element (Haegeman and Greco, 2018; Lybaert et al., 2019).

(1) Met zulk weer je kunt niet veel doen.
(SAND sentence 359)
with such weather you can NEG much do
"With such weather, you cannot do much."

[1] In this paper, we clearly distinguish between *dialect* and *accent,* regarding accent as "restricted to phonological and especially phonetic differences, such as the quality of vowel sounds" (Boberg et al., 2018, p. 4).

[2] Debate is possible on the exact delineation of the SDDs (cf. Taeldeman and Hinskens, 2013); for reasons of comparability, the corpus project described in this paper will span the same geographical area as other major dialectological projects (cf. e.g., Van Keymeulen et al., 2019).

[3] In the near future, we hope to expand the corpus, also by collaborating with the Meertens Institute, to include recordings from their dialect database (the so-called *Nederlandse Dialectenbank,* 'Dutch Dialect Bank').

[4] The standard language is mainly based on the sociolect of the middle class in the cities of the provinces North- and South-Holland (see Willemyns, 2003 for a historical background).

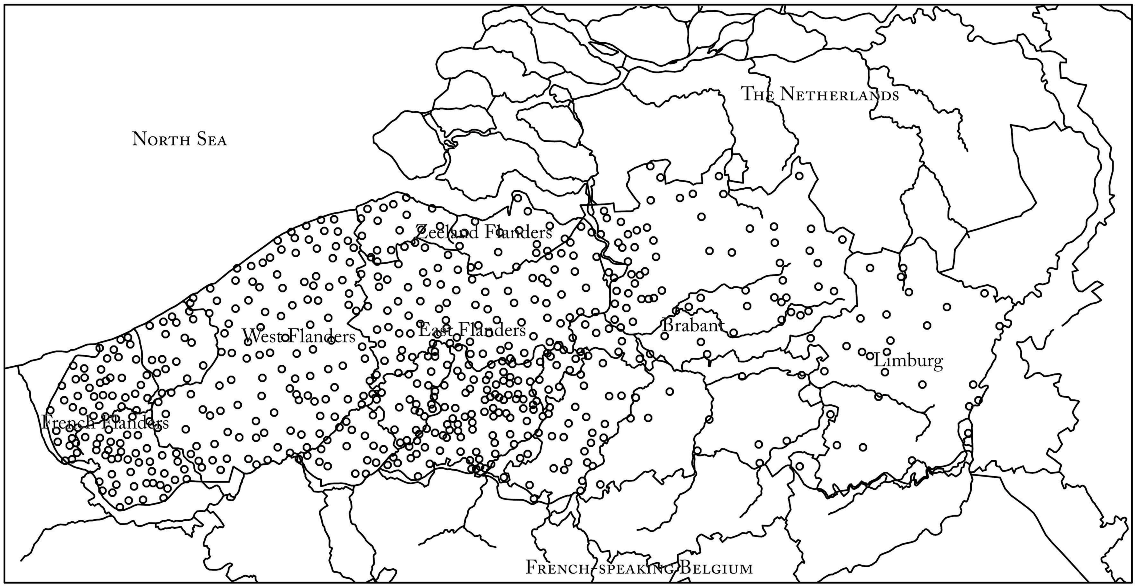


FIGURE 1 | Regional spread of the dialect recordings of the collection "Stemmen uit het Verleden."

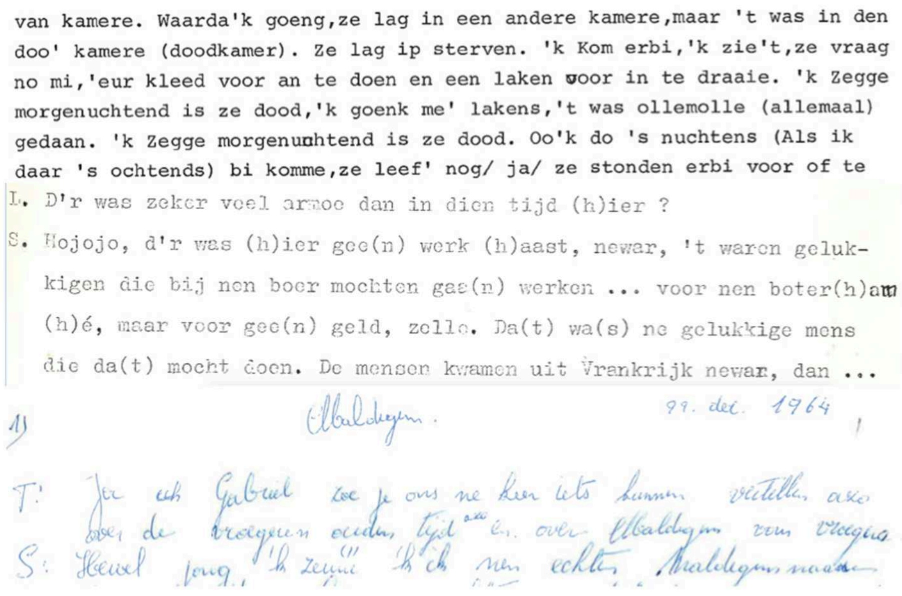
Data such as (1) are underreported in the *Syntactische Atlas van de Nederlandse Dialecten* (SAND, 2005/2008), as many types of these so-called V3 constructions are only realized in very specific pragmatic contexts (e.g., to indicate that something comes as a surprise) and are hence difficult to elicit in a survey. Indeed, for several locations, the SAND fieldworkers observe that even though rejected by the informants, the pattern is attested in their spontaneous speech, which the notes of the fieldworkers acknowledge.[5] This is only one example of a phenomenon that would benefit from being studied on the basis of a corpus of spontaneously spoken dialect (complementary to survey data analysis).

In the 1960s and 1970s, dialectologists at Ghent University made 783 tape recordings of 45 min on average (in total about 700 h) in 550 locations (cf. **Figure 1**) in the Dutch-speaking provinces in Belgium, Zeeland Flanders (Netherlands), and French Flanders (France). Their goal was to build a corpus for dialect research. The recorded speakers—often practitioners of an occupation considered vanishing or 'lost' at the time of recording—are born in the late nineteenth and early twentieth centuries (the oldest speaker was born in 1871) and are almost always monolingual dialect speakers, and because most of the speakers have received only minimal formal education, their speech is hardly influenced by the Dutch standard language. The speakers were generally interviewed by a fellow villager in the local dialect to avoid adaptation to the language of the interviewer. The topics of the conversations were free; in contrast to, for instance, the interviews for the SAND or the *Syntax hessischer Dialekte* (SyHD, Fleischer et al., 2015), the aim was not to elicit specific linguistic constructions. In general, the speakers narrate about their life, profession, and the sociocultural changes they witnessed during their lifetime. This makes the material, which has become known under the name *Stemmen uit het Verleden* ('Voices from the past'), valuable not only for linguistic purposes, but also for (oral–) historical and cultural–historical reasons.

The collection of dialect recordings constitutes a valuable data source both for large-scale phonological, morphological, lexical, and syntactic research and for the study of specific phenomena that are mainly restricted to spontaneous speech, and which therefore resist elicitation. Because the speakers recorded were born around the turn of the twentieth century, and hence acquired language about 100–120 years ago, these recordings already represent a historical stage of the language. Additionally, the tapes contain accounts of oral history that may provide valuable information on, e.g., the events around the World Wars. Moreover, the recordings constitute a treasure trove of cultural heritage, such as lost professions and customs.

The accessibility of the recorded dialect data is undeniably invaluable for linguistic and historical research. However, the vast collection of data can currently hardly be used for linguistic or historical research, as the material is not digitally searchable for word forms (allowing one to make concordances of keywords in context), let alone for syntactic patterns and constructions. Thanks to various projects such as *Stemmen uit het verleden* ("Voices from the past"; see www.dialectloket.be), the tapes have been digitized and safeguarded for posterity. Nevertheless, various hurdles have to be overcome to make the material fully accessible for researchers. Firstly, only 318 of the 783 recordings have been transcribed. These transcriptions were generally made in the 1960s and 1970s by students writing a dissertation on dialect syntax. Secondly, the transcriptions that exist (i) are

[5] Cf. the 'comment' section in the data from the surveys on http://www.meertens.knaw.nl/sand/zoeken/lijst_met_plaatsen.php



van kamere. Waarda'k goeng,ze lag in een andere kamere,maar 't was in den
doo' kamere (doodkamer). Ze lag ip sterven. 'k Kom erbi,'k zie't,ze vraag
no mi,'eur kleed voor an te doen en een laken voor in te draaie. 'k Zegge
morgenuchtend is ze dood,'k goenk me' lakens,'t was ollemolle (allemaal)
gedaan. 'k Zegge morgenuchtend is ze dood. Oo'k do 's nuchtens (Als ik
daar 's ochtends) bi komme,ze leef' nog/ ja/ ze stonden erbi voor of te

L. D'r was zeker veel armoe dan in dien tijd (h)ier ?

S. Hojojo, d'r was (h)ier gee(n) werk (h)aast, newar, 't waren geluk-
kigen die bij nen boer mochten gae(n) werken ... voor nen boter(h)am
(h)é, maar voor gee(n) geld, zollo. Da(t) wa(s) ne gelukkige mens
die da(t) mocht doen. De mensen kwamen uit Vrankrijk newar, dan ...

1) Maldegem. 19. dec. 1964

T: Ja ook Gabriel zoe je ons ne keer iets kunnen vertellen over
de vroegeren ouden tijd en over Maldegem van vroeger
S: Heeuwel joeng, 'k zeune 'k ben nen echten Maldegemnaar

FIGURE 2 | Excerpts from existing transcriptions for recordings in Torhout, Wichelen, and Maldegem, respectively.

only available electronically in the form of scans (i.e., image files) of the original typewritten or even handwritten texts, (ii) often contain many mistakes, (iii) are not time-aligned to the audio (cf. infra), and (iv) are heterogeneous in the way the dialect has been transcribed. This heterogeneity can be attributed to the fact that there is hardly a writing tradition in the dialect—dialects have been passed on orally from generation to generation—and that only a brief transcription guideline was provided. **Figure 2** illustrates the heterogeneity by means of three excerpts from existing transcriptions, two in typoscript and one in handwriting. Whereas in the first and third excerpt, non-standard Dutch vocalism is rendered in a kind of 'eye dialect' (e.g., in excerpt 1: *ip* and *ollemolle* instead of standard Dutch *op* 'on' and *allemaal* 'all'; in excerpt 3: *zeune* instead of standard Dutch *zijn* 'be'), non-standard vocalism has been standardized in the transcription from Wichelen. The dialectal vowel in *gaan* ('go'), which is pronounced as [ɔ˙] in the recording from Wichelen, is for instance not rendered in the transcription. A similar heterogeneity can be seen in the way the deletion of initial or final consonants is marked: in the first excerpt, apostrophes are used (e.g., *me*' for standard Dutch *met* 'with'), while in the second, the deleted consonant is reconstructed between brackets [e.g., *da(t)* for standard Dutch *dat* ('that')]. These are only some examples of the heterogeneity in the existing dialect transcriptions. Bearing in mind the currently rapidly advancing dialect loss across Flanders (Vandekerckhove, 2009; Ghyselen and Van Keymeulen, 2014), there is a real risk that soon there will not be any speakers able to understand and hence to (help) transcribe them. In order to make this unique collection of dialect data present at Ghent University accessible for fundamental research, their transcription and linguistic annotation is therefore of high priority. Achieving these two goals is the core of the project *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* (GCND, Spoken Corpus of the SDDs).

PROTOCOL REQUIREMENTS

Transcribing is a process of data reduction: some elements of the speech signal are visualized in the transcript; others are ignored. A transcript is hence always a research construct (Jenks, 2011, p. 11), the result of numerous decisions on which elements to graphically render and which not. As dialect corpus research requires transcripts to be "as reliable, faithful, and internally consistent as possible" (Szmrecsanyi and Anderwald, 2018, p. 302), it is of utmost importance that a detailed protocol is developed, which ensures that all transcribers take the same or rather similar decisions in the data reduction process.

In devising a protocol, it is vital to keep the purposes of the collection in mind, which in the case of the GCND are diverse. On the one hand, the transcripts in our corpus have to cater to the needs of linguists interested in the diverse aspects of the dialect system. The main purpose of the GCND is to provide a database for both corpus-based and corpus-driven research (Biber, 2009) on the syntax of the Dutch dialects, as syntactic patterns—especially optional constructions (Cornips and Poletto, 2005, p. 955)—are known to be especially difficult to study via elicitation. However, the corpus should ideally of course also allow morphological, lexical, and phonological/phonetic research, e.g., dialectometric research measuring the phonetic distance between dialects (cf. Heeringa, 2004; Nerbonne and Heeringa, 2010). In this context, (i) a high transcription accuracy and consistency is needed and (ii) transcribers cannot simply standardize non-standard words, pronunciations, or constructions, as this is exactly what dialectologists are interested in. On the other hand, the transcripts should also be accessible for historians, ethnologists, or laymen interested in the content of the tapes. For this reason, the texts should also be readable to those not thoroughly familiar with phonetic alphabets or all the specificities of the local dialects. A further consideration in choosing a transcription protocol is that the transcripts should allow NLP tools to automatically annotate the texts with POS tags[6] and syntactic parsing information and that such tools are typically trained on standard language resources and hence benefit from transcriptions close to standard language norms.

To allow phonetic or phonological research, the dialect transcriptions should preserve as much phonetic detail as possible. However, manual phonetic transcriptions are—more than other types of transcriptions—very sensitive to transcriber effects, and hence pose a problem for transcript consistency. Bailey et al. (2005) discuss how, even after careful phonetic training of transcribers, the phonetic transcriptions needed for the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) and the *Linguistic Atlas of the Gulf States* (LAGS) were clearly subject to transcriber effects, due to "(1) conceptual differences regarding the phonetic status of particular sounds (e.g., offglides of diphthongs) and how they should be transcribed, (2) normative differences regarding the phonetic values of particular symbols, and (3) changing scribal practices as transcribers discover the importance of phonetic details that they had previously overlooked" (Bailey et al., 2005, 3). Phonetic transcription is also much more time- and hence also budget-consuming than orthographic transcription. In the context of the GCND, for which a large number of transcribers collaborate on the transcription of 700 h of very diverse spoken dialects, it was quickly decided that manual phonetic transcription was simply not feasible. As an alternative, time and effort was invested to link all layers (or tiers) of annotations to time codes in the audio, thereby ensuring that researchers interested in phonetic detail can easily consult the passages of the audio relevant for their research purposes and provide phonetic annotations themselves. Below (see section 'Forced alignment for automatic segmentation and phonetic transcription'), we also investigate the possibility of automatic phonetic transcription via forced alignment (FA).

Leaving the option of manual phonetic transcription aside, the question arises how non-standard pronunciations, lexical items, and syntactic structures—in the absence of writing norms for Dutch dialects—should be rendered in the Latin alphabet. In addressing this question, a difficult balance has to be struck between faithfulness to the original dialect on the one hand and regularization to guarantee consistency, searchability, and accessibility for non-linguists on the other hand. Given that interoperability and sustainability/reusability are important requirements in the year 2019 when collecting and annotating data (cf. the philosophy of CLARIN, the European Research Infrastructure for Language Resources and Technology; De Jong et al., 2018), it is interesting to consider how this balancing act has been performed in earlier variationist research.

BUILDING ON EXISTING STANDARDS

The approaches chosen in existing projects transcribing non-standard or dialect speech range from almost entirely using standard orthography to layered transcriptions differentiating phonological variation and standard orthography. The COSER[7] and PRESEEA[8] corpora of dialectal and spoken Spanish, for instance, use one layer of orthographic transcription, which also represents a number of divergences from the standard language orthographically. In COSER transcripts, non-standard stress positions and omissions and additions of phonological segments are systematically rendered orthographically. The FOLK corpus of spoken (near-standard) German (Schmidt, 2016)[9] and the CORDIAL-SIN corpus of Portuguese dialects[10] in principle transcribe orthographically in one layer except for some individual words. FOLK is transcribed in a modified orthography ('eye dialect,' cf. Schmidt, 2016, p. 119) following the GAT2-standard (Selting et al., 2009, also used in the research project *Deutsch in Österreich*[11]). The corpus, however, also provides normalized forms for divergent items as word-level tags to the original transcription. This normalized transcription is the input to further NLP processing. The transcription in CORDIAL-SIN[12] uses the standard orthography even in the case of regionally divergent phonology, except in cases that are considered as potentially relevant for future (morpho)syntactic analysis. In such cases, divergent phonetic realizations, contractions, and truncations are marked in the same layer by stating the divergent form and the standard form next to each other, e.g., deu-{PH|lɨ=lhe} for (standard Portuguese) *deu-lhe* '(he) gave him' for phonetic variation in the pronunciation of the clitic *lhe*, or {IP|'pɛɾɐ=espera} for the truncation of the initial part of the word *espera* 'wait.IMPV.' For further NLP processing (morphological tagging and syntactic parsing), only the normalized form is used, which is produced automatically from the original transcription by replacing, e.g., {IP|'pɛɾɐ=espera} by *espera*. This normalized form is stored in a separate file (ASCII and .pdf). The last possibility, fully transcribing in two layers, with one layer representing the original dialect and one 'translating' the dialect into standard orthography, is not used by existing spoken language corpora as far as we are aware. This is presumably due to a large degree of overlap between the produced dialect strings and the standard language. Such overlap is however

[6] These tags indicate the grammatical category (e.g., noun, adverb, or preposition) of each word in the text.

[7] COSER = *Corpus Oral y Sonoro del Español Rural* ('Audible Corpus of Spoken Rural Spanish'). http://corpusrural.es

[8] PRESEEA = *Proyecto para el Estudio Sociolingüístico del Español de Españay de América* ('Project for the Sociolinguistic Study of Spanish from Spain and America'). http://preseea.linguas.net/Corpus.aspx

[9] FOLK = *Forschungs- und Lehrkorpus Gesprochenes Deutsch* ('Research and Teaching Corpus of Spoken German'). http://agd.ids-mannheim.de/folk.shtml

[10] CORDIAL-SIN = *Corpus Dialectal para o Estudo da Sintaxe* ('Syntax-oriented Corpus of Portuguese Dialects'). https://clul.ulisboa.pt/en/recurso/cordial-sin-syntax-oriented-corpus-portuguese-dialects

[11] https://dioe.at/

[12] https://clul.ulisboa.pt/sites/default/files/inline-files/manual_normas.pdf

somewhat problematic for the SDDs, as there is a high degree of phonological, lexical, morphological, and syntactic divergence between dialects and standards, which complicates a procedure marking all forms diverging from the standard language with individual tags.

In the Dutch language area, there are no digital corpora of spontaneously spoken dialect yet. The large-scale Corpus of Spoken Dutch (*Corpus Gesproken Nederlands* or *CGN*, Oostdijk, 2000)—containing approximately 9 million words—focuses on (intended) standard language, and hence its transcription protocol is not geared toward dialect research.[13] There is however a rich tradition of dialect study in the Dutch language area (cf. Goossens and Van Keymeulen, 2006) and, as such, there are already conventions for dialect orthography to be built upon. For the GCND protocol, Barbiers and Vanden Wyngaerd (2001) was taken as point of departure, who describe the transcription guidelines used for the *Syntactic Atlas of Dutch Dialects*.[14] For this SAND project, transcriptions were made of questionnaires—asking for the judgment and/or translation of some 150 test sentences—conducted orally (fieldwork and telephone) between 2000 and 2005 in about 300 locations across The Netherlands, Belgium, and a small part of north-west France. The protocol was devised with syntactic purposes in mind, and hence opts for strong standardization of non-standard pronunciation in content words, whereas non-standard functional elements (inflection, pronouns, articles, auxiliaries, etc.) and syntactic structures (word order, double negation, and extra complementizers) are transcribed as closely to the dialect as possible. For the GCND protocol, a comparable approach was adopted:

- Phonological variations of content words that also exist in the standard language are spelled according to official standard language orthography (as established by the Dutch Language Union in the *Woordenlijst Nederlandse Taal*).[15] If a speaker for instance pronounces the standard language word *steen* ([steːn] 'stone') with a diphthong (e.g., [sti·ən]), we write *steen*; for reasons of intertranscriber consistency and readability, these non-standard pronunciations are not rendered in some kind of 'eye dialect' (we do not write *stieën*).
- Content words that do not have an equivalent in the standard language are written down following the principles of Standard Dutch spelling as closely as possible. The word [lɑtəsto·ərs] for instance ('roll-down shutters') is written down as *lattestoors*. Non-standard words are *not* translated into a standard Dutch alternative (such as *rolluik*), as (i) these non-standard lexemes are of interest to dialectologists and (ii) the precise translation of these dialect words is often open to debate. If the non-standard words have already been included in an existing dialect dictionary (e.g., www.e-wvd.be for the Flemish dialects, www.e-wbd.nl for the Brabantic dialects, and www.e-wld.nl for the Limburgian dialect), transcribers adopt the dictionary spelling. To guarantee transcriber consistency, a logbook of non-standard lexemes and their spelling is shared among transcribers.
- Function words (inflection, adpositions, auxiliaries, determiners, negation particles, conjunctions, and pronouns) are transcribed as close to the dialect as possible, with an orthographic rendering of deletions and insertions of consonants (cf. Moreno et al., 2016 on the Spanish COSER corpus). If a speaker pronounces *wat* ('what') without final [t], the deletion is also written down (*wa*). Vocalic changes with functional value (e.g., *vuut* 'foot.pl,' standard Dutch *voeten* 'feet,' with the umlaut marking the plural) are also transcribed, following standard Dutch orthographic rules as accurately as possible. Regular changes in the vocalism [e.g., the pronunciation of standard Dutch [aː] as [ɔː] in for instance *maar* ('but')] are however not transcribed, but rendered in standard Dutch spelling, as trying to consider all these phenomena would compromise the consistency among transcribers.
- Non-standard clitics [e.g., *tkind* for standard Dutch *het kind* ('the child')] are written down as clusters of elements, using hashtags to mark—intuitively—the different elements part of the cluster (e.g., *t#kind*). This 'hashtag analysis' is not a fixed fact, but has the status of a 'first guess' (cf. Barbiers and Vanden Wyngaerd, 2001, p. 6).
- Non-standard syntactic constructions (e.g., with subject duplication or alternative word orders) are transcribed as close to the dialect as possible.

To cater to the needs of non-linguists intending to search the content of the tapes and to facilitate the functioning of NLP tools—which are mainly trained on standard language data—an extra transcription layer is added in the GCND corpus, a layer in which function words are standardized (*gunder* or *gider*, e.g., are written down as *jullie* 'you [plural]') and clitics are separated into their component parts (e.g., *t#kind* is written down as *het kind* 'the child'). In this standardized layer, non-standard lexemes and non-standard constructions are preserved, as it is often unclear what the standard language equivalents for these words and constructions should be. All layers of transcription are time-aligned to the audio using ELAN (Max Planck Institute for Psycholinguistics, cf. Brugman and Russel, 2004).[16] Example sentence (2) showcases the different principles outlined in GCND protocol.

[13] In the CGN project, transcribers were instructed to use words from a predefined lexicon, which contained (in principle) all Standard Dutch words and also a number of commonly occurring reduced forms. Dialect words or constructions not occurring in Standard Dutch (and hence also not in the CGN lexicon) are marked with the label '*d', whereas standard Dutch words pronounced in a "heavily dialectal way" get the label '*z'. Such an approach is not feasible for a dialect corpus project, as about any word, word form, or construction would have to be marked by either '*d' or '*z'.

[14] These data are now freely available online via http://www.meertens.knaw.nl/sand/zoeken/. With the MIMORE tool (http://www.meertens.knaw.nl/mimore/search/), the researcher can search in the *Dynamic Syntactic Atlas of the Dutch Dialects* combined with two other databases (*Diversity in Dutch DP* and the *Goeman, Taeldeman, van Reenen Project*) with a common online search engine. The search results can be visualized on geographic maps and exported for statistical analysis.

[15] Can be consulted via www.woordenlijst.org.

[16] https://tla.mpi.nl/tools/tla-tools/elan/

(2) Recording N72_Ieper.

[ʌptlɑətstəvɑneːɦənɛntwintəhksɪniːrkɔmnwærkŋɪndəfilɑtyːrə]					*IPA-transcription*
op t#laatste van negenentwintig	k#zijn	ier	komen werken	in de filature.	*Layer 1*
op het laatste van negenentwintig	ik ben	hier	komen werken	in de filature	*Layer 2*
At the end of twenty nine	I am	here	come work	in the filature.	*Gloss*
"At the end of twenty nine I started working here in the filature."					*Translation*

The first layer in (2) stays close to the original dialect, orthographically rendering:

- non-standard words, here the dialect word *filature* for standard Dutch *spinnerij* ('filature'),
- non-standard morphology (e.g., *zijn* for standard Dutch *ben* 'am'),
- clitics (e.g., *t#laatste* for standard Dutch *het laatste* 'the end'),
- the insertion or deletion of consonants in function words (e.g., *ier* for standard Dutch *hier* 'here'), and
- non-standard syntax [cf. word order ADVERBIAL *(op t#laatste van negenentwintig)* + SUBJECT *(k)* + CONJUGATED VERB *(zijn)* instead of Standard Dutch ADVERBIAL + CONJUGATED VERB + SUBJECT in main clauses].

Non-standard variations of content words that also exist in the standard language are however standardized. We, for instance, write *negen* 'nine' and not *nehen*, even though the speaker clearly laryngalizes the fricative [ɤ]. In the second layer, the non-standard lexeme *filature* and non-standard syntactic constructions (lack of inversion after the adverbial phrase) are preserved, but clitics are written down as clusters of elements (*t#laatste* > *het laatste*), deleted consonants are 'restored' (*ier* > *hier*), and the morphology is standardized (*k#zijn* > *ik ben*).

SPEECH TECHNOLOGY TO THE RESCUE?

The transcription procedure outlined above is—when performed manually—very time-intensive and therefore expensive. Transcription speeds for our data range from 67 s/h for a beginning transcriber to 120 s/h for an experienced one. The question arises whether speech technology can speed up the process. In what follows, we review a number of methods that can potentially accelerate the transcription and/or alignment process: automatic speech recognition (ASR, section Automatic speech recognition), respeaking, and forced alignment (FA).

Automatic Speech Recognition

In the last few decades, significant headway has been made in ASR. ASR analyzes the sound spectrum of the input speech and tries to determine—on the basis of a language or even dialect specific *acoustic model*—which phonemes could correspond to the input spectra. An acoustic model contains statistical representations for each phoneme in a language, created from a set of audio recordings and their corresponding transcripts. Next, the obtained set of phonemes is used to estimate via a (dedicated) *language model* the words that could have been spoken. A language model is a statistical model that represents the probabilities of words and phrases in a specific language. The result of this estimation process is a set of words with their start time, duration, and recognition probability. Modern ASR engines like the KALDI and Google recognizers can recognize 256K different words.

ASR has many applications. It is for instance increasingly used for spoken document retrieval, as illustrated by the FAME! Project (Frisian Audio Mining Enterprise). This project developed an ASR system for Frisian–Dutch code-switching speech, as extracted from the archives of a local broadcaster. The goal of the system was to allow automatically retrieving relevant items from a large collection of news broadcasts, in response to user-specified text queries (Yilmaz et al., 2018, p. 12). Similarly, Van Den Heuvel et al. (2012) report applying ASR to disclose—via keyword retrieval—250 interviews with veterans of Dutch conflicts and military missions. ASR also has applications in reporting. Kawahara (2012) discusses the development of a speaker-independent ASR system for transcribing plenary and committee meetings of the Japanese Parliament. This system is said to consistently produce accuracy levels of at least 85%. The automatically generated transcripts are then further processed by parliamentary reporters. The usefulness of ASR for reporting purposes, however, strongly depends on the language under study. An innovation project carried out in Flanders in 2017–2018 led to the conclusion that the state of speech-to-text technology for Dutch was insufficient at the time to be useful for reporting debates of the Flemish Parliament, as it did not increase, but rather reduced reporting efficiency.[17]

In linguistic research, ASR remains little used for full automatic transcription. There are, however, examples of successful application. Michaud et al. (2018) for instance describe how ASR advanced the study of Yongning Na, a Sino-Tibetan language of Southwest China. Of the 14 h of speech the authors recorded during fieldwork, 5.5 h (both narratives and morphotonology elicitation sessions) were transcribed by hand. Subsequently, an ASR transcription tool was trained on these transcribed materials, in order to perform phoneme recognition on the remaining untranscribed audio files. The error rate of the resulting transcriptions proved low, about 17%. According to the authors, the automatic transcriptions reduced the manual effort required for creating transcripts and allowed new insights that might not have been gained by the linguist alone.

Via user-friendly interfaces building on neural network models (e.g., Cloud Speech-to-Text by Google), even computational laymen can now attempt to convert audio to text automatically. A quick test in Google Cloud Speech-to-Text on 129 words of intended Standard Dutch, as spoken by a highly educated West Flemish speaker in a standard language test (cf. Ghyselen, 2016) yields a fine Word Error

[17] http://innovatieveoverheidsopdrachten.be/projecten/spraaktechnologie-voor-verslaggeving-vlaams-parlement

Rate (WER)[18] of only 7%. As ASR systems can also add time codes in the transcription—useful to align the text to the original audio—ASR offers interesting opportunities for speech corpus building.

However, many dialects—such as the Southern Dutch ones—must be considered 'low resource languages,' i.e., languages for which few tools and/or resources are available. This constitutes a major challenge for the application of ASR. While acoustic and language models for Netherlandic Dutch and Belgian Dutch have been developed (cf. https://www.spraak.org and https://spraaktechnologie.org), these were mainly trained on standard language and on regionally colored speech, which is much closer to the standard norm than the dialects in our data collection. Generally, tools trained on standard language underperform on non-standard data. While the intended standard Dutch sample of the highly educated West Flemish speaker discussed above yielded a WER of only 7% in Google Cloud Speech-to-Text (cf. Ghyselen, 2016), the WER increased to 66% in a test using 164 words of a spontaneous interview by the same West Flemish speaker. This is high, considering that the language used in the interview is not fully fledged dialect, but only diverges in some pronunciation features from the official standard language (especially h-dropping and t-dropping) and that the recording quality was high. Note that the option 'Netherlandic Dutch' had to be used, as 'Belgian Dutch' was not available. The low-resource problem is—as can be expected—only exacerbated with dialect data. Tested on a dialect recording from the *Voices of the Past* collection[19], Google Cloud Speech-to-Text obtains a WER of 90%. A comparison of the reference transcription in (3) with the automatic transcription in Google Cloud Speech-to-Text (option: Netherlandic Dutch) in (4) illustrates how ASR is at present not helpful as a tool to speed up the transcription process in the GCND project.

(3) k#e vijf jaar in Tourcoing ewrocht in e fabrieke. van negen... uh van drieëntwintig tot negenentwintig. en in ne... op t#laatste van negenentwintig k#zijn ier komen werken in de filature. vierendertig jaar. en k#e moeten twee jaar eerder mijn pensioen nemen. omda#k epakt waren aan mijn harte. en ezo k#zijn nu gepensioneerd. k#zijn nu tweeënzestig nu nieuwjaar. twee dagen voor nieuwjaar zij#k tweeënzestig. ja en k#zijn al elf jaar mijn man kwijt wi. awel ja#k. ja k#e maar een zoone.

I have worked for five years in Tourcoing in a factory. from nine… uh from twenty-three to twenty-nine. and in… in the end from twenty-nine I have come here to work in the filature. thirty-four years. and I have had to retire two years earlier. because I had heart problems. and as such I am retired now. I'll turn sixty-two at new year. two days before New Year I am sixty-two. yes and I have lost my husband for eleven years already. yes I have. yes. I have only one son.

(4) fabrieken van 23 tot 29 van 29,34 jaar omdat tweedehands
factories from 23 to 29 from 29.34 years because second-hand

The ASR tool of the BASWebServices of the Ludwig Maximilian University of Munich[20] performed equally poorly (WER = 95%), with the following output:

(5) koel je nou vooral ten fabrieken van van drieëntwintig tot negenentwintig van negenentwintig vierendertig jaar heb ik haar mond nu twee jaar om daskapan kwamen bij maar dat is ook zien in een gepassioneerd twee dan voor een nieuwe hadden ja die arme man with we elkaar Morrison

cool you now especially at factories of of twenty-three to twenty-nine from twenty-nine thirty-four years have I her mouth now two years for daskapan[21] came at but this is also see in a passionate two then for a new had yes that poor man with we each other Morrison.

Of course, ASR tools (including the acoustic and language models) can be adjusted/retrained on new data to cater to the needs of dialectologists, but currently, no suitable tools exist. Furthermore, the retraining of such tools typically requires large amounts of already transcribed text from all dialects to be efficient.

In deliberating the usefulness of ASR investments (e.g., developing dialect-specific acoustic and language models) in a dialect corpus project, there are different factors to consider. A first one is the sound quality of the audio collection: recordings with background noise, much overlapping speech and/or a large variance in recording settings (distance from microphone etc.), present a bigger challenge for ASR systems. Michaud et al. (2018, p. 396) point out that the high audio quality of their recordings of Yongning Na speech is an important part of the reason why the automatic transcription yielded good results. The authors stress that for low-resource languages, it is highly important that the pronunciation is clear and the audio signal is clean. In the case of our dataset, the recordings were made in the 1960s and 1970s in 550 locations (often private homes of dialect speakers, with barking dogs, ticking clocks, or vehicles passing by as background noise) with reel-to-reel tape recorders and often multiple speakers per recording. The acoustic properties as such differ from recording to recording, which implies serious challenges for ASR systems.

Secondly, the performance of ASR tools strongly depends on the degree of linguistic differentiation between the dialects and standard language. As explained above, the SDD systems diverge significantly phonologically, morphologically, syntactically, and lexically from the Dutch standard language, which explains why tools developed for Netherlandic (Standard) Dutch perform

[18] The WER is "the edit distance between a reference word sequence and its automatic transcription, normalized by the length of the reference word sequence" (McCowan et al., 2005, 2): WER = $(S+D+I)/N_r$, with N_r as the total words in the reference transcription, S as the number of substituted words in the automatic transcription, D as the number of words from the reference transcription deleted in the automatic transcription, and I as the number of words inserted in the automatic transcription not appearing in the reference. See McCowan et al. (2005) for a critical discussion.

[19] A short excerpt (89 words) was selected from the dialect recording of Ieper (West Flanders).

[20] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface, language selected: "Dutch (Netherlands) – conversation."

[21] Words such as *daskapan* also make no sense in Dutch.

so poorly on SDD recordings. It is not easy to extend the existing tools for non-standard speech, as this requires a significantly large training set of transcribed dialect, which is not available for the Dutch dialects. In the last few years, the dictionaries of the Flemish, Brabantic, and Limburgian dialects have become available online (cf. e-wvd.be, e-wbd.nl, and e-wld.nl, respectively), which offers opportunities for ASR systems, but the keywords in these dictionaries are ‘standardized’[22] — given the lack of orthographic norms for the dialects (cf. section ‘Toward a corpus of Southern Dutch Dialects’)—and as such, the ASR systems will need enough training data to link the acoustic realization of non-standard words to the keywords in the dictionaries. An important issue is also the diversity among dialects: global tools simultaneously trained on many dialects have been reported to “fail to generalize well for any of them,” as a consequence of which state-of-the-art speech recognition systems, including that of Google, prefer building a different recognizer per “dialect” (Elfeky et al., 2018, p. 2). In the case of the SDDs, the diversity is so large—with four big dialect areas that are internally also very diverse morphologically, phonologically, syntactically, and lexically—that multiple recognition systems should be built, implying serious time and financial investments.

A third factor to bear in mind is the goals the ASR transcripts have to serve, as this determines the transcript accuracy needed. For example: the FAME! Project already introduced above obtained WERs ranging from 32 to 33% (Yilmaz et al., 2018, p. 18–19), which is a satisfying result when the goal of the transcripts is to make the broadcast archive more searchable content-wise. Ordelman et al. (2007, p. 214) mention a WER of 50% as baseline for spoken document retrieval. However, when the goal of a project is to facilitate linguistic research, a higher transcription accuracy is needed. If the researcher has to correct 1 out of 2 words manually after implementing ASR, he might as well not lose time (and money) on ASR and transcribe the speech manually from the beginning. A tough question to answer is what ballpark area WERs have to be in for ASR (or another speech technological tool) to become a viable option in linguistic research, e.g., to provide a first draft. Human transcribers are said to have error rates ranging between 3 and 10%, depending on the type of input speech and the time spent on the transcription (Stolcke and Droppo, 2017, p. 137–138). In the context of the GCND project, a comparison of student transcriptions for four recordings of four different dialect areas with the final equivalent as corrected by both an older speaker of the recorded dialect and the project coordinator yields an average WER of 3% (lowest = 0.4%, highest = 6.4%). This WER is difficult—not to say impossible—to equal with ASR (at least when it concerns non-standard speech), but there is still the option of first creating a draft transcript using ASR and then manually correcting it. Ranchal et al. (2013) report the results of such an approach to transcribe lectures taught in English. They obtained—after voice profile training—WERs of 22% for the automated first transcript. The manual correction of these automated transcripts is said to take 4 h per hour of lecture audio (Ranchal et al., 2013, p. 306–307), which still is a lot, given that the researchers also invested time in the ASR development and voice profile training. It hence seems logical to assume that with WERs of over 30%, it is time and budget friendlier to transcribe the recordings manually from the start.

Considering the issues discussed above in the context of the GCND, the decision was made not to invest in ASR development, given (i) the very diverse acoustic properties of the recordings, (ii) the current lack of training data, (iii) the diversity among the SDDs and the large distance between these dialects and the standard Dutch varieties for which ASR tools have already been developed, and (iv) the high transcription accuracy needed for the further linguistic annotation and analysis of the dialect data. Of course, once the corpus is available, the transcripts can be used to train new dialectal/regiolectal recognizers of Dutch.

Respeaking

As discussed above, quality requirements for dialect transcriptions can at present often not be met by state-of the-art ASR technology. There are however other alternatives to a purely manual transcription approach, combining human skill, and speech technology. Sperber et al. (2013), for instance, suggest *respeaking* to provide a good trade-off between transcription quality and cost. In *respeaking,* a speaker repeats and records the speech of the original speaker using a speech recognition system. *Respeaking* is assumed to be faster than typing, and allows circumventing some of the problems in ‘pure’ ASR approaches, as the respeaker’s voice can be recorded in a strictly controlled setting (cf. sound quality problem discussed above) and the ASR system can be trained or adapted to the voice of the respeaker.

Respeaking is nowadays often used to (i) subtitle live broadcasts (cf. Imai et al., 2002; van Waes et al., 2013), typically when there is no script available (Romero-Fresco, 2011) or (ii) to lower the cost of speech transcription via crowd-powered speech transcription platforms (cf. Vashistha et al., 2017). Of course, as respeaking partly builds on ASR tools, it is also sensitive to errors. Therefore, an editor or the respeaker often manually corrects the initial draft transcription (van Waes et al., 2013, p. 18) or ASR transcripts of the same audio respoken by multiple respeakers are compared and combined (Vashistha et al., 2017).

Respeaking also has applications in linguistic research and, in fact, in a dialect corpus project somewhat similar to the GCND. For the Spanish COSER corpus, a respeaker approach is adopted to build a parsed corpus of European Spanish dialects (Rufino Morales, 2019). One respeaker from Granada, who understands most peninsular Spanish dialects well, has been trained to respeak interviews made between 1990 and now.

By way of trial, one of the authors of this paper—a variationist linguist and native speaker of the West Flemish dialect—respoke the excerpt in Example (3), standardizing non-standard vocalism. The resulting audio was then fed into the ASR tool of the BASWebServices of the Ludwig Maximilian University of

[22] Adjusted to standard Dutch spelling systems and regularizing dialectal pronunciation features. The West Flemish word [ɦəbøːrneːɦə] is for instance written down as *gebuurnege* (Standard Dutch *buurvrouw* ‘female neighbor’), in which the West Flemish laryngalisation of standard Dutch /ɣ/ and the West Flemish realization of West Germanic û before /r/ as [øˑ] are standardized and written down as <g> and <uu> (cf. standard Dutch [ɣ] and [yˑ]), respectively.

Munich.[23] The WER of the resulting transcript [see (6), with, for the sake of convenience, also a repetition of the manual reference transcription in (7)]—34%—is remarkably lower than the one obtained by applying ASR on the original audio (95%). Thirty-four percent is still high—as already discussed at the end of the previous section a WER of this size still requires too much manual correction to be useful—but it might be seen as a sign that with the necessary training and technical optimization, respeaking could be a valuable technique in the transcription process.

(6) k intern qua verankerd in de fabriek van een van mevrouw drieëntwintig tot negenentwintig en in nee op het laatste zijn hier komen werken in de file vierendertig jaar en k moeten twee jaar eerder mijn pensioen nemen onderdak pakt waren aan mijn hart en zo ik zijn nu gepensioneerd zijn nu tweeënzestig nu nieuwjaar twee dagen voor het nieuwe jaar zei tweeënzestig ja en ik zijn al elf jaar mijn man kwijt wil ja ik ga maar één

I internal qua anchored in the factory of a of madam twenty-three to twenty-nine and in no at the end am here come work in the traffic-jam thirty-four year and I have to two year earlier my retirement am now sixty-two now new year two days before the new year said sixty-two yes and I am already eleven year my husband lost want yes I go but one

(7) k#e vijf jaar in Tourcoing ewrocht in e fabrieke. van negen... uh van drieëntwintig tot negenentwintig. en in ne... op t#laatste van negenentwintig k#zijn ier komen werken in de filature. vierendertig jaar. en k#e moeten twee jaar eerder mijn pensioen nemen. omda#k epakt waren aan mijn harte. en ezo k#zijn nu gepensioneerd. k#zijn nu tweeënzestig nu nieuwjaar. twee dagen voor nieuwjaar zij#k tweeënzestig. ja en k#zijn al elf jaar mijn man kwijt wi. awel ja#k. ja k#e maar een zoone.

I have worked for five years in Tourcoing in a factory. From nine… uh from twenty-three to twenty-nine. and in… in the end from twenty-nine I have come here to work in the filature. thirty-four years. and I have had to retire two years earlier. because I had heart problems. and as such I am retired now. I'll turn sixty-two at new year. two days before New Year I am sixty-two. yes and I have lost my husband for eleven years already. yes I have. yes. I have only one son.

There are, however, a number of issues to bear in mind, in particular with respect to projects like the current one. Firstly, the respeaker must understand the dialect(s) well. In the case of the COSER corpus, the respeaker from Granada is able to cover a lot of the Iberian Peninsula, but in other language communities and also when it concerns older recordings, affected less by dialect leveling, such wide intelligibility is everything but self-evident (cf. Boberg et al., 2018, p. 5 on mutual intelligibility of dialects and clines of linguistic similarity). The (southern) Dutch dialects for instance, as they are recorded in the *Stemmen uit het Verleden* collection, display significant linguistic differences between each other, as well as with the standard language, on which the tools are trained. As stated earlier, these differences also concern such typological traits as word order and inflectional morphology. Cliticization and pronoun doubling are cases in point. In (8), five clitics form a cluster that behaves like one phonological word. In order to transcribe such a sequence adequately using respeaking, separate pronunciation on the part of the respeaker is required. This would require the respeaker to parse such clitic clusters in real time.

(8) Recording H68_Loppem
k#en#e#k#ik nooit niet gezien
ik en heb ik ik nooit niets gezien
I NEG have I I never nothing seen
"I have never seen anything."

The historical SDDs in the collection already show significant typological differences already within a short geographical distance. As it is highly unlikely that one could find a single respeaker capable of understanding all these dialects, multiple respeakers [e.g., (at least) one per dialect region] would have to be trained for the GCND. This implies that the ASR software would also have to be trained for multiple speakers. Evidently, the time and money needed to (a) train these respeakers, (b) (re)train the ASR systems, and (c) correct the draft transcripts is not sufficiently compensated by the gain in time respeaking is said to have over typing. Secondly, respeaking requires quickness of response to the original audio (Romero-Fresco, 2011). In the case of the GCND audio collection, which actually represents historical speech, transcribers often consult dialect dictionaries and studies on local customs and folklore to determine what the dialect speakers in the recordings are talking about. This of course complicates the respeaking process. Thirdly, respeaking is also sensitive to some of the problems encountered when discussing ASR (cf. section Automatic speech recognition), e.g., the training data needed to adjust the ASR system. The advantage of respeaking is that the respeaker can standardize dialectal pronunciations of standard language words, but of course (i) such standardization requires a serious cognitive effort and (ii) the respeaking system still has to be able to handle dialectal lexemes (especially when the goal is to build a dialect corpus). At the same time, certain morphological, syntactic, and lexical phenomena should in fact not be standardized, as argued above. For all these reasons, the decision was made not to use respeaking in the GCND project.

Forced Alignment for Automatic Segmentation and Phonetic Transcription

Another alternative to 'pure' ASR that combines speech technology with human effort is FA, the process of aligning speech (audio) with text (the written representation of the recorded speech). FA requires transcriptions as input (made manually or automatically), and as such does not clear the transcription hurdle. It does, however, allow (i) automatically creating phonetic transcriptions on the basis of orthographic

[23] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface, language selected: "Dutch (Netherlands) – conversation."

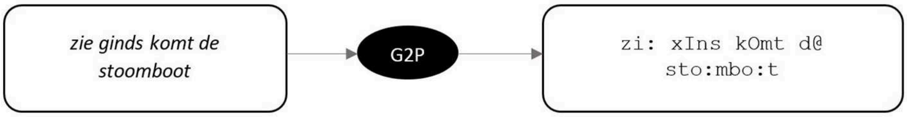


FIGURE 3 | Grapheme-to-phoneme conversion of the Dutch sentence *Zie ginds komt de stoomboot* ('see the steamboat over there').

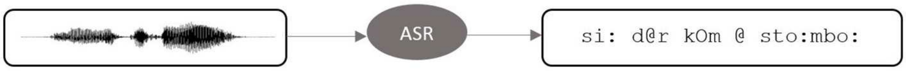


FIGURE 4 | Automatic speech recognition (from audio to transcription).

ones and (ii) automatically aligning the text transcription to the audio on a word or phoneme level (the latter is also called phonetic alignment).

In FA, the input text is parsed into a chain of words and subsequently passed to a grapheme-to-phoneme (G2P) algorithm (cf. **Figure 3**), which results in a string of phonemic symbols.[24] As a rule, this happens via the canonical transcriptions of the words in the text, i.e., the way in which—according to some predefined standard (either specifying the pronunciation rules of a language or combining a lexical pronunciation dictionary with fallback to the rule-based system)—the words ought to be pronounced. More advanced G2P algorithms also take into account phonetic processes that occur when combining certain words (e.g., assimilation) or pronunciation variants that may occur in spontaneous speech [cf. WORDVAR in the Munich Automatic Segmentation (MAUS) system, (Schiel, 1999)], but nonetheless, the phonetic rendering is always based on how the words in the text are expected to be pronounced on the basis of a defined standard or system, not on how the speaker has actually pronounced these words.

Parallel to the G2P conversion, the speech signal is transcribed phonetically by means of ASR (cf. **Figure 4** and the earlier section on automatic speech recognition). In the case of the example in **Figures 3**, **4**, the pronunciation of the speaker, as 'decoded' by ASR, does not entirely match the canonical transcription made on the basis of the input text (e.g., with devoicing of the /z/ in the word *zie* in the speech signal).

A next step consists of aligning the outputs of both G2P and ASR (the actual FA), attempting to match the two sequences as 'efficient' as possible. In **Table 1**, gray cells represent phonemes where there is a match between the two outputs, yellow cells involve substitutions and red cells indicate that an 'expected' sound is not detected in the actual speech signal.

As the speech recognizer determines begin and end times for each of the detected sounds, it is possible to calculate the begin and end times of the words, even when the 'dictionary' pronunciation does not (entirely) match the actual pronunciation. As such, the text transcriptions can be linked to the audio on a phoneme and word level, allowing researchers interested in the pronunciation of specific words or sounds to find these more easily in a speech corpus and to export the relevant portions of the audio efficiently into speech analysis software (such as Boersma and Weenink, 2011). However, the accuracy of the time codes does decrease inversely proportional to the differences between the norm pronunciation and the actual pronunciation.

Some FA applications also allow automatic phonetic transcription. The Munich Automatic Segmentation system (Schiel, 1999) for instance generates, on the basis of the canonical phonetic transcription of an orthographic transcription fed into the system, an acyclic-directed graph of all probable pronunciation variants of the input utterance, along with the predictor probability of these variants. Subsequently, the graph and the speech wave are "passed to a standard Viterbi alignment procedure that computes the best combined probability of acoustical score and predictor probability, in other words, finds the most likely path through the graph" (Schiel, 1999, p. 2). As such, a (broad) phonetic transcript is created that combines information from (i) the speech signal (the actual speech), (ii) an orthographic transcription, and (iii) specified knowledge about the pronunciation of a certain language.

FA has many applications in linguistic research. The corpus of spoken Dutch (CGN) for instance applied FA not only to align the speech signal at word level to the orthographic transcription but also to automatically generate broad phonetic transcriptions of about 900 h of recorded speech on the basis of orthographic transcriptions. Goddijn and Binnenpoorte (2003) report error rates ranging from 15% for spontaneous speech to 6% for read speech and conclude that automatic phonetic transcription on the basis of orthographic transcripts is the best approach for their spoken (near-)standard Dutch data, in combination with manual correction. The inverse procedure is also possible: creating an orthographic transcription departing from a phonetic one. In the Nordic Dialect Corpus for instance, all Norwegian dialects and some Swedish ones were first transcribed phonetically, and subsequently, the phonetic transcriptions were translated to orthographic ones via a semi-automatic dialect transliterator developed for the project (Johannessen et al., 2009). Of course,

[24] Phonemes are generally written down in SAMPA (Speech Assessment Methods Phonetic Alphabet), which is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA).

TABLE 1 | Forced alignment of G2P and ASR output.

Input text	*zie*		*ginds*				*komt*				*de*		*stoomboot*						
G2P output	z	i:	x	I	n	s	k	O	m	t	d	@	s	t	o:	m	b	o:	t
ASR output	s	i:	d	@	r		k	O	m			@	s	t	o:	m	b	o:	
Speech	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T8*	*T9*	*T10*	*T11*	*T12*	*T13*	*T14*	*T15*	*T16*	*T17*	*T18*	*T19*

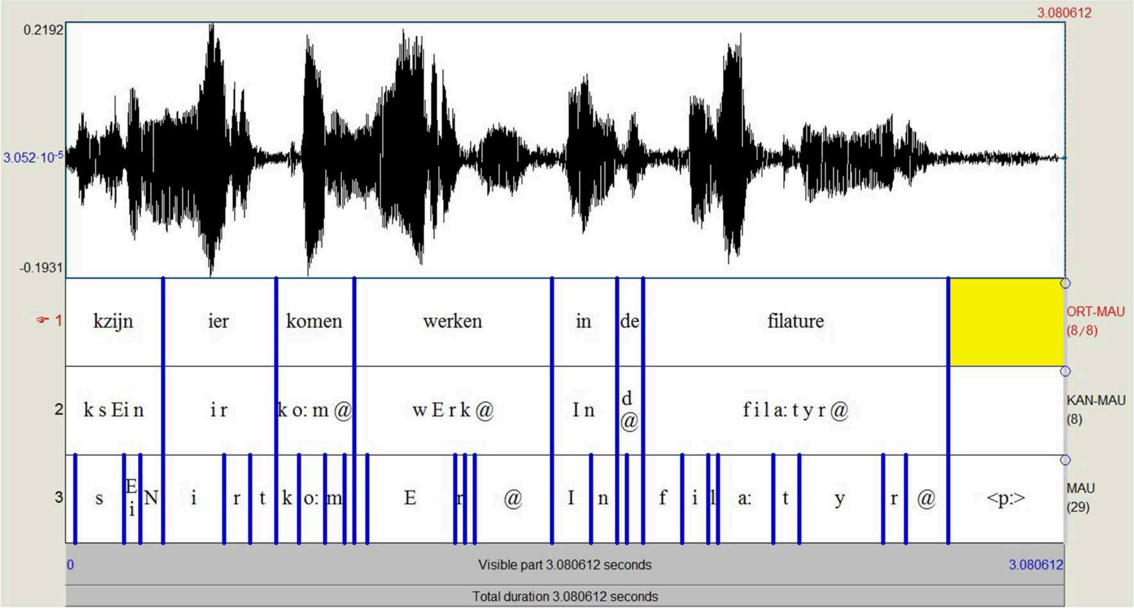


FIGURE 5 | WebMAUS output for the dialect sentence *kzijn ier komen werken in de filature* ('I have started to work here in the filature').

manual phonetic transcription is more time-consuming than manual orthographic transcription. Another application of FA can be found in the automatic extraction of variables for phonetic analysis (cf. Evanini et al., 2009 and Rosenfelder et al., 2014 on the FAVE automated vowel extraction program and Reddy and Stanford, 2015 on DARLA, which automatically generates transcriptions with ASR and extracts vowels using FAVE).

Figure 5 shows the output of a FA test using the BASWebServices of the Ludwig Maximilian University of Munich (Schiel, 1999; Kisler et al., 2017).[25] Their WebMAUS-module segments an audio file into SAMPA phonetic segments given an orthographic transcription. We fed the dialect sentence *kzijn ier komen werken in de filature* [cf. Example (2) above] with the corresponding audio into WebMAUS, selecting as language 'Dutch (BE).' The first layer ('ORT-MAU') shows the original orthographic transcription (following the project protocol). The second layer ('KAN-MAU') represents the canonical phonetic transcriptions created by the G2P algorithm on the basis of 'Dutch_BE' as specified language, and the third layer ('MAU') shows the automatic phonetic transcription, representing the best combined probability of acoustical score and predictor probability (cf. supra).

We subsequently tested the accuracy of the WebMAUS aligner on a slightly longer stretch of West Flemish dialect speech [Example (3)]. The word boundaries determined by the forced aligner are remarkably accurate: 81% of the 90 words are accurately delimited, notwithstanding the fact that the pronunciation of the speaker deviates clearly from the standard Dutch pronunciation of the words used. The phonetic transcription and delineation of phonemes (cf. layer 3 in **Figure 5**) are a bit less accurate, but still good. We obtain a phoneme error rate[26] of 28%, which is not perfect—it certainly is not good enough to use for phonetic research without manual correction—but it is also not disastrous, especially considering the absence of acoustic, and language models for the SDDs. The automatic phonetic segmentation and transcription could be improved, by either (i) feeding phonetic transcriptions into the system, making it possible to skip the G2P procedure or—less ideal, but more feasible—(ii) departing from an orthographic transcription that renders more of the pronunciation peculiarities in the text than our current

[25] Cf. https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

[26] The Phoneme Error Rate was calculated in a similar way to the word error rate: (S+D+I)/Nr, with Nr as the total number of phonemes in the reference transcription, S as the number of substituted phonemes in the automatic transcription, D as the number of phonemes from the reference transcription deleted in the automatic transcription, and I as the number of phonemes inserted in the automatic transcription not appearing in the reference.

transcriptions do ('eye dialect transcription'). Concerning (i), we already indicated that phonetic transcriptions are too time-consuming and too prone to intertranscriber inconsistencies. Concerning (ii), fixed rules in the dialect (e.g., <ij> in orthography should be pronounced as [i] in many West Flemish dialects) can be specified to automatically add pronunciation information in existing transcriptions. In the case of the SDDs, however, 'dialect rules' often depend from place to place and are in many cases also lexically diffuse (meaning that a rule applies to some words, but not to others). It is hence difficult to list 'dialect rules' that apply to all words with a specific orthography in all SDDs. As an alternative, we deliberated adjusting the transcription protocol in such way that the first layer of the transcription (closest to the dialect) would be more of an 'eye dialect' rendering than was the case in the SAND protocol from which we departed. This can be seen as a middle course between full phonetic transcription and a more standardized orthographic transcription, which might improve the automatically generated phonetic transcriptions. We therefore retested the WebMAUS FA on Example (3), now with an orthographic transcription that marked more dialectal pronunciations. In this new transcription, we, for instance, wrote *zin* and *kwit* instead of *zijn* ('to be') and *kwijt* ('lost') to indicate that the old West Germanic î is realized as a monophthong [i] before non-labiodentals in many West Flemish dialects. The new transcription also marked (i) schwa-deletions (e.g., by writing *moetn* instead of *moeten* 'have to'), (ii) h-deletions (*erte* instead of *harte* 'heart'), (iii) the shortening of [aː] to [ɑ] (*latste* instead of *laatste* 'last'), (iv) the palatalization of [oː] in certain words (e.g., *zeune* instead of *zoone* 'son'), (v) the velarization of [aː] to [ɔː] (e.g., *joar* instead of *jaar* 'year'), and (vi) the realization of an intervocalic [j] in words such as *drie(j)ëntwintig* ('twenty-three'). Fed into the FA system, this adapted orthographic transcription did not improve the word segmentation success (now 79% of the 91 words were correctly delineated), but it did cause a decrease of 5% in the phoneme error rate (resulting in an error rate of 23%).

Our test results indicate that FA can be very useful for dialect corpus building. In the context of the GCND corpus, we decided to apply FA for word-level segmentation. This word-level segmentation is interesting as it allows searching for and extracting the pronunciation of individual words in the corpus, useful in, for instance, lexical, and phonological projects. Phonological/phonetic research is not the primary goal of the corpus project, but all the same the intention is to make the corpus as multi-usable as possible. Word-level segmentation also allows a detailed alignment of word-level annotations (such as POS tags) to the audio. Given the low error rates the aligner obtained with our data, it seems possible to apply word-level alignment without much manual correction. Manual correction is however clearly needed when FA is applied for automatic phonetic transcription. FA can certainly speed up the transcription process by providing a rough first transcription as a starting point, but to make this useful for phonetic research, a serious time investment is still needed. For the GCND project, the decision was therefore made not to invest in FA for phonetic transcription. Phoneticians interested in the corpus can, however, of course apply FA themselves to create phonetic transcriptions. We also decided not to alter the original transcription protocol in the direction of a more 'eye dialectal' rendering of non-standard vocalism, as the improvement this rendering brought for FA was in our opinion too small to compensate for the extra complexity eye dialect renderings add to the manual transcription process. Hence, the decision was made to stick with the original transcription protocol, as this guaranteed more consistency among transcribers.

TRANSCRIPTION PROCEDURE IN THE GCND

After weighing the advantages and disadvantages of existing speech technological tools for the transcription of dialectal speech, the decision was made to manually transcribe the dialect recordings of the 'Voices from the Past' collection in two layers, each aligned to the audio at sentence level using the software package ELAN (Max Planck Institute for Psycholinguistics, cf. Brugman and Russel, 2004). This manual transcription is very time-consuming—with transcription speeds for our data ranging from 67 s/h for a beginning transcriber to 120 s/h for an experienced transcriber—but it is at the moment still the most efficient option, as ASR has much difficulties handling the SDDs and as such yields transcriptions with error rates that are too high to be useful for linguistic research. Speech technology, and more specifically FA, can however be helpful to automatically refine the rough manual alignment of the transcription to the audio (which happens at sentence or clause level) to a word-level alignment, facilitating phonetic research.

A difficult question in the GCND project was what to do with the existing 318 transcriptions, which—as mentioned in section Toward a corpus of Southern Dutch Dialects —are currently only available in the form of scans (i.e., image files) of the original typewritten or even handwritten texts. It is of course possible to use optical character recognition (OCR) on these image files and have a forced aligner align the resultant text files to the audio file, but the problem remains that the transcriptions are very heterogeneous in the way the dialect has been transcribed orthographically (cf. **Figure 2**) and that the original transcriptions often contain many mistakes. Also considering that the OCR and FA procedures would cause extra mistakes (given the diversity in input image files, cf. **Figure 2**), manual editing would still be necessary, adjusting the texts to the new protocol, adding a second layer of transcription, and correcting mistakes of both the transcriber, the OCR and the forced aligner, which raises the question whether it is not more time-efficient to make a new (manual) transcription from scratch, using the original transcription as resource to speed up the transcription process. The decision was made to not invest time in optimizing and executing OCR and FA procedures, as manual labor was necessary anyhow.

Of vital importance when working with human transcribers is that a detailed, yet workable transcription protocol is developed and that sufficient training is provided. For the GCND project, five student-transcribers[27] tested a first version of the protocol described in the section 'Building on existing standards'. They were asked to keep a log of problems they encountered during transcribing, which was subsequently discussed during weekly group meetings with the project leaders. During this test phase, the protocol was refined and elaborated with examples. A next group of 15 student-transcribers was hired and trained to work with the new protocol. To guarantee transcription accuracy and consistency, all students received (i) a group demo of the software and the protocol, (ii) online training materials, (iii) personalized feedback on their initial transcriptions (random samples were corrected by the project supervisors), and (iv) access to a shared 'problem database,' where dubious cases could be registered and the project supervisors subsequently offered advice on how to transcribe the problematic utterance in line with the protocol.

Of course, human transcribers are also not infallible. To guarantee the quality of the transcriptions, a crowd-sourcing network has been established in which volunteers check the transcriptions made by student-transcribers. These volunteers especially focus on speech fragments marked with the code "???" by the transcribers. The ??? code indicates passages that the student-transcribers did not understand, either because of gaps in their dialect proficiency or because of limited familiarity with the speech topic (e.g., when the interviewee talks about farming techniques or barrel making).[28] Contrary to the student-transcribers, most volunteers acquired the traditional dialect as a first language. They generally also have more life experience—the majority of volunteers have retired—and are hence usually more tuned in to the subject matter than the student-transcribers. The volunteers check the accuracy of the transcriptions on paper or text files exported from ELAN; their corrections and additions are evaluated and adjusted in ELAN by a project worker fully acquainted with the protocol and the software. As already mentioned in the section on Automatic speech recognition, comparison of initial student transcriptions with the final, corrected equivalents for four recordings of four different dialect areas yields an average WER of 2.93%, which, in comparison with the WERs of ASR tools, is very low and argues in favor of manual transcription.

CONCLUSIONS AND RECOMMENDATIONS FOR PRACTICE

There are at present many speech technological tools available that can speed up the transcription of spontaneous speech, such as ASR, respeaking, and FA, but dialects—at least when defined in the 'traditional' sense as a regionally determined language varieties which differ at multiple structural levels from other dialects and the 'overarching' standard language—still constitute a major challenge. For the transcription of the dialect audio collection available at Ghent University (*Stemmen uit het Verleden* 'Voices from the past'), the choice was made to use speech technology only for the word-level segmentation of the audio files, as the transcription itself could not be sped up by ASR tools. This decision is however not necessarily also appropriate for other dialect corpus projects. In deliberating the usefulness of speech technological tools for a dialect corpus project, the following questions have to be considered:

- **What is the sound quality of the recordings?** If the recording quality is high (with a high-quality external microphone, little background noise, or overlapping speech and a similar distance to the microphone for all speakers), speech technological tools should be considered. Recordings of poorer quality, however, with more interference and more heterogeneous speech, still pose a major challenge for speech technological tools such as ASR, particularly in the absence of suitable models. This problem can, when the conditions discussed below are favorable, be circumvented using respeaking. As respeaking combines ASR with human 'labor'—a respeaker repeats and records the speech of the original speaker using a speech recognition system—poor audio quality or heterogeneity of the original speech can be set right in the first step of the respeaking process.
- **Which resources are available for the dialect(s) under study?** Application of ASR can be considered if a pronunciation dictionary for the dialect(s) has been developed, or—even better—if acoustic and language models are available for the dialect(s) and/or overarching standard language. When pronunciation dictionaries or acoustic or language models are only available for the standard language, and not for the dialect(s) under study, the usefulness of speech technological tools strongly depends on the way in which standard and dialect(s) differ.
- **What is the degree of linguistic differentiation between the dialects in the corpus and the standard language?** If the distance between the dialects is large and no straightforward rules can be formulated about the correspondences between these dialects (e.g., sound X in dialect A always corresponds to sound Y in dialect B or in the standard language, cf. Rys and Van Keymeulen, 2009), multiple recognition systems have to be built for ASR (or tools integrating ASR, such as FA), implying serious time and financial investments. If the distances, however, are small, or systematic correspondence rules can be listed for the differences between the dialects or between the dialects and the standard language, it can be considered to develop dialect-specific acoustic and language models for ASR tools. Linguistic differentiation is also an important criterion when considering the usefulness of respeaking. If the dialects under study are mutually intelligible, one respeaker can be trained to handle the whole dataset. If the dialects are not or only partially mutually intelligible, respeaking poses a bigger challenge.
- **Which goals do the transcripts have to serve?** If the main goal is to make recordings searchable in terms of content, a moderate transcription accuracy (with WERs up to 50%) is often perfectly acceptable, and such accuracy can be achieved using speech technology. If the transcripts, however, have to

[27] All transcribers working in the pilot project were students in linguistics at Ghent University.

[28] Fragments that were difficult to transcribe for acoustic reasons (e.g., recording quality, background noise or overlapping speech) are marked with a different code ("xxx").

serve as input for linguistic research, a higher transcription accuracy is needed. In that case, the researcher has to weigh the advantages of a procedure consisting of ASR [with or without respeaker(s)] and subsequent manual correction against those of manually transcribing the recordings from the beginning. Also, it should be considered which type of linguistic research the transcripts have to facilitate. If the focus is mainly on syntax, lexicon, or morphology, an orthographic transcription of the original audio suffices and word-level alignment of the audio to the transcription is perfect. Such word-level alignment can be perfectly achieved—in case one ultimately decides not to transcribe with ASR from scratch—with the help of FA. If phonetic research is intended, FA can also automatically generate phonetic transcriptions on the basis of orthographic ones. Manual correction is still needed, but the broad phonetic transcription created by FA can speed up the phonetic transcription process (cf. above).

Only when all of these questions have been addressed is it possible to decide whether or not to invest in ASR development for dialect transcription. In case the deliberation militates in favor of manual transcription, it is important that a detailed protocol is developed and tested in interaction with multiple transcribers and that sufficient attention is paid to the training of transcribers, with the necessary opportunities for feedback.

In all probability, significant headway will in the next few years be made in the automatic recognition of non-standard speech. While the interests of computational linguists and dialectologists might diverge at some points—as dialect shift and leveling processes progress, the dialects in the 'Voices from the past' collection for instance increasingly represent a historical stage of the language, which is greatly interesting for linguists modeling theories on language variation and change, but might appeal less to computational linguists training speech recognizers to handle everyday speech—cooperation between dialectologists and ASR specialists is undoubtedly fruitful. Speech recordings transcribed and annotated manually by dialectologists are useful training materials for computational linguists, even when the dialects represent the language of only a fraction of a speech community. In diglossic communities for instance (Auer, 2005), where a continuum of intermediate varieties has developed between the traditional dialects and the official standard language (e.g., in Dutch-speaking Belgium or Germany), intermediate varieties are generally marked by a combination of dialect and standard language variants. In such contexts, a speech recognizer that can handle both local dialects *and* standard language can handle a large part of the sociolinguistic repertoire. To be continued…

DATA AVAILABILITY STATEMENT

The dialect recordings discussed in this contribution can be consulted freely on www.dialectloket.be.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Faculty of Arts and Humanities at Ghent University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JV, AB, and A-SG devised the project. AB was in charge of overall direction and planning. A-SG developed the transcription protocol with help from JV. MF contributed transcriptions to the project, was—together with A-SG and AB—responsible for the training of the student-transcribers, and computed the WERs of the manual transcription procedure. A-SG performed the tests discussed in the sections on speech technological tools, under the guidance of AH with input from all authors. A-SG wrote the manuscript.

FUNDING

The pilot project on which this paper is based was funded by a small research grant by the Research Foundation Flanders (FWO) to AB [Grant No. 1.5.310.18N], by a postdoctoral fellowship (FWO) to MF [Grant No. 12P7919N], and by grants by the provinces of West Flanders (KIOSK-projectID 29575), East Flanders (KIOSK-projectID 29311), and Zeeland (ref. of confirmation letter: 19004695) to Variaties vzw. This paper was published with the support of the Belgian University Foundation.

ACKNOWLEDGMENTS

We kindly thank (i) our student-transcribers for their diligent transcription work, (ii) all volunteers correcting the student transcriptions, (iii) Lien Hellebaut, for her conscientious coordinating and correction work, and (iv) Veronique De Tier, for her feedback and help.

REFERENCES

Auer, P. (2005). "Europe's sociolinguistic unity, or: a typology of European dialect/standard constellations," in *Perspectives on Variation*, eds N. Delbecque, J. van der Auwera, and D. Geeraerts (Berlin; New York, NY: Mouton De Gruyter), 7–42.

Bailey, G., Tillery, J., and Andres, C. (2005). Some effects of transcribers on data in dialectology. *Am. Speech* 80, 3–21. doi: 10.1215/00031283-80-1-3

Barbiers, S., and Vanden Wyngaerd, G. (2001). *Transcriptieprotocol voor de Syntactische Atlas van de Nederlandse Dialecten.* Amsterdam: Meertens Instituut.

Biber, D. (2009). "Corpus-based and corpus-driven analyses of language variation and use," in *The Oxford Handbook of Linguistic Analysis,* eds B. Heine and H. Narrog (Oxford: Oxford University Press), 159–191.

Boberg, C., Nerbonne, J., and Watt, D. (2018). "Introduction," in *The Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Oxford: John Wiley & Sons), 1–15. doi: 10.1002/97811188276280

Boersma, P., and Weenink, D. (2011). *Praat: Doing Phonetics by Computer [Computer Program]*. Version 5.2.46. Retrieved from: http://www.praat.org/ (accessed September 30, 2011).

Breitbarth, A., and Haegeman, L. (2014). The distribution of preverbal *en* in (West) Flemish: syntactic and interpretive properties. *Lingua* 147, 69–86. doi: 10.1016/j.2013.11.001

Brugman, H., and Russel, A. (2004). "Annotating multimedia/multi-modal resources with ELAN," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (Lisbon), 2065–2068.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Cornips, L., and Poletto, C. (2005). On standardising syntactic elicitation techniques (part 1). *Lingua* 115, 939–957. doi: 10.1016/j.2003.11.004

De Jong, F., Maegaard, B., De Smedt, K., Fiser, D., and Van Uytvanck, D. (2018). "CLARIN: towards FAIR and responsible data science using language resources," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018) (Miyazaki), 3259–3264.

De Schutter, G. (2009). De ontwikkeling van negatiepatronen met niet-negatieve onbepaalde kwantoren in de midden- en zuid-brabantse dialecten. *Taal Tongval* 61, 1–20. doi: 10.5117/TET2009.2.DESC

De Vogelaer, G. (2008). *De Nederlandse en Friese subjectsmarkeerders: geografie, typologie en diachronie*. Gent: Koninklijke academie voor Nederlandse taal-en letterkunde.

Elfeky, M. G., Moreno, P., and Soto, V. (2018). Multi-dialectical languages effect on speech recognition. Too much choice can hurt. *Proc. Comput. Sci.* 128, 1–8. doi: 10.1016/j.2018.03.001

Evanini, K., Isard, S., and Liberman, M. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. *Proc. Interspeech* 2009, 1655–1658.

FAND = Goossens, J., Taeldeman, J., and Verleyen, G. (1998: deel I; 2000: deel II + III); De Wulf, C., Goossens, J., and Taeldeman, J. (2005: deel IV). *Fonologische Atlas van de Nederlandse Dialecten.* Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Fleischer, J., Lenz, A. N., and Weiß, H. (2015). "Syntax hessischer Dialekte (SyHD)," in *Regionale Variation des Deutschen. Projekte und Perspektiven*, eds R. Kehrein, A. Lameli, and S. Rabanus (Berlin: De Gruyter), 261–287.

Ghyselen, A-S. (2016). *Verticale structuur en dynamiek van het gesproken Nederlands in Vlaanderen: een empirische studie in Ieper. Gent en Antwerpen*, Ph.D. Ghent University.

Ghyselen, A-S., and Van Keymeulen, J. (2014). Dialectcompetentie en functionaliteit van het dialect in Vlaanderen anno 2013. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 130, 17–139.

Goddijn, S., and Binnenpoorte, D. (2003). "Assessing manually corrected broad phonetic transcriptions in the spoken dutch corpus," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona), 1361–1364.

Goossens, J., and Van Keymeulen, J. (2006). "De geschiedenis van de Nederlandse dialectstudie," in *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 80, 37–97.

Haegeman, L., and Greco, C. (2018). West flemish V3 and the interaction of syntax and discourse. *J. Comp. Ger. Linguist.* 21, 1–56. doi: 10.1007/s10828-018-9093-9

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Rijksuniversiteit Groningen.

Imai, T., Matsui, A., Homma, S., Kobayakawa, T., Onoe, K., Sato, S., et al. (2002). "Speech recognition with a re-speak method for subtitling live broadcasts," in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002.* Denver, Colorado, USA.

Jenks, C. J. (2011). *Transcribing Talk and Interaction: Issues in the Representation of Communication Data*. Amsterdam, PA: John Benjamins.

Johannessen, J. B., Priestley, J. J., Hagen, K., Åfarli, T. A., and Vangsnes, Ø. A. (2009). "The nordic dialect corpus–an advanced research tool," in *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA* 2009), 73–80.

Jonas, D., Whitman, J., and Garrett, A. (eds.). (2011). *Grammatical Change: Origins, Nature, Outcomes*. Oxford: Oxford University Press.

Kawahara, T. (2012). "Transcription system using automatic speech recognition for the Japanese parliament (Diet)," in *Proceedings of the 24th Innovative Applications of Artificial Intelligence 3* (Toronto, ON), 2224–2228.

Kemmer, S., and Barlow, M. (2000). "Introduction: a usage-based conception of language," in *Usage-based Models of Language Use*, eds M. Barlow and S. Kemmer (Stanford: CSLI Publications), 7–28.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347. doi: 10.1016/j.2017.01.005

Lybaert, C., De Clerck, B., Saelens, J., and De cuypere L. (2019). A corpus-based analysis of V2 Variation in West Flemish and French Flemish Dialects. *J. Ger. Linguist.* 31, 43–100. doi: 10.1017/S1470542718000028

MAND=De Schutter, G., van den Berg, B., Goeman, T., and De Jong, T. (2005; deel I: Meervoudsvorming bij zelfstandige naamwoorden, vorming van verkleinwoorden, geslacht bij zelfstandig naamwoord, bijvoeglijk naamwoord en bezittelijk voornaamwoord); Goeman, T., Van Oostendorp, M., van Reenen, P., Koornwinder, O., and van den Berg, B. (2009; deel II: comparatief en superlatief, pronomina, werkwoorden presens en preteritum, participia en werkwoordstamalternaties). *Morfologische Atlas van de Nederlandse Dialecten.* Amsterdam: Amsterdam University Press.

Mathieu, E., and Truswell, R. (eds.). (2017). *Micro-Change and Macro-change in Diachronic Syntax.* Oxford: Oxford University Press.

McCowan, I., Moore, D. C., Dines, J., Gatica-perez, D., Flynn, M., Wellner, P., et al. (2005). *On the Use of Information Retrieval Measures for Speech Recognition Evalation.* Martigny: IDIAP Research Institute.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: experiments with na data and the persephone toolkit. *Lang. Doc. Conserv.* 12, 393–429.

Moreno, C. B., Pueyo, J., and Fernández-Ordóñez, I. (2016). "Creating and designing a corpus of rural Spanish," in *KONVENS 2016, 20 de septiembre de 2016*, eds S. Dipper, F. Neubarth, and H. Zinsmeister (Bochum: Rühr-Universität Bochum), 78–83.

Nerbonne, J., and Heeringa, W. (2010). "Measuring dialect differences," in *Language and Space. An International Handbook of Linguistic Variation. Theories and Methods*, eds P. Auer and J. E. Schmidt (Berlin: Mouton De Gruyter), 550–566. doi: 10.1515/9783110220278.550

Oostdijk, N. (2000). "The Spoken Dutch Corpus. Overview and first Evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, eds M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (Athens: LREC), 887–894.

Ordelman, R. J. F., de Jong, F. M.G., and Van Leeuwen, D.A. (2007). "Speech Indexing," in *Multimedia Retrieval,* eds H. Blanken, A. P. de Vries, H. E. Blok, and L. Feng (Berlin; Heidelberg; New York, NY: Springer), 199–224.

Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, P. J., et al. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Trans. Learn. Technol.* 6, 299–311. doi: 10.1109/TLT.2013.21

Reddy, S., and Stanford, J. N. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguist. Vanguard* 1, 15–28. doi: 10.1515/lingvan-2015-0002

Romero-Fresco, P. (2011). *Subtitling through Speech Recognition: Respeaking*. Manchester/Kinderhook: St. Jerome Publishing.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard H., et al. (2014). *FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3.* Software https://doi.org/10.5281/zenodo.9846

Rufino Morales, M. R. (2019). "El rehablado off-line para potenciar la transcripción de un corpus oral en español," in *Talk at I Congreso Internacional de Lingüística Digital-CILiDi'19*, Granada.

Rys, K., and Van Keymeulen, J. (2009). Intersystemic correspondence rules and headwords in Dutch dialect lexicography. *Int. J. Lexicogr.* 22, 129–150. doi: 10.1093/ijl/ecp008

SAND = Barbiers, S., Hans, B., Gunther De, V., Magda, D., and van der Ham, M. (2005; deel I: Pronomina, Congruentie en Vooropplaatsing); Barbiers, S., Bennis, H., De Vogelaer, G., Van der Auwera, J., and van der Ham, M. (2008; deel II: Werk-woordsclusters en negatie). *Syntactische Atlas van de Nederlandse Dialecten.* Amsterdam: Amsterdam University Press.

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Procedures of the ICPhS* 1999, 607–610.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the research and teaching corpus of spoken german. *Int. J. Corpus Linguist.* 21, 396–418. doi: 10.1075/ijcl.21.3.05sch

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., et al. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung* 10, 353–402.

Sperber, M., Neubig, G., Fügen, C., Nakamura, S., and Waibel, A. (2013). Efficient speech transcription through respeaking. *Interspeech* 2013, 1087–1089.

Stolcke, A., and Droppo, J. (2017). "Comparing human and machine errors in conversational speech transcription," in *Interspeech, Vol. 2017* (Stockholm), 137–141.

Swanenberg, J., and van Hout, R. (2013). "Recent developments in the mid southern dialects," in *Language and Space. An International Handbook of Linguistic Variation. Dutch*, eds F. Hinskens and J. Taeldeman (Berlin/Boston: Walter de Gruyter), 319–335.

Szmrecsanyi, B., and Anderwald, L. (2018). "Corpus-based approaches to dialect study," in *The Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Oxford: John Wiley & Sons), 300–313.

Taeldeman, J., and De Wulf, C. (2010). "Niet-suffigale eind-sjwa's in de Vlaamse dialecten," in *Voor Magda. Artikelen voor Magda Devos bij haar afscheid van de Universiteit Gent*, eds J. De Caluwe and J. Van Keymeulen (Gent: Academia Press), 591–611.

Taeldeman, J., and Hinskens, F. L. M. P. (2013). "The classification of the dialects of Dutch," in *Language and Space. An International Handbook of Linguistic Variation. Dutch*, eds F. Hinskens and J. Taeldeman (Berlin/Boston: De Gruyter Mouton), 129–142.

Trudgill, P. (1999). *The Dialects of England. 2nd Edn*, Oxford/Malden: Blackwell.

Van Den Heuvel, H., Sanders, E., Rutten, R., Scagliola, S., and Paula, W. (2012). An oral history annotation tool for INTER-VIEWs. *LREC* 215–218.

Van Keymeulen, J., de Tier, V., Vandenberghe, R., and Chambers, S. (2019). The dictionary of the Southern Dutch dialects. Designing a virtual research environment for digital lexicological research. *Dialectologia et Geolinguistica. J. Int. Soc. Dialectol. Geolinguist.* 8, 93–115.

van Waes, L., Mariëlle, L., and Remael A. (2013). Live subtitling with speech recognition. causes and consequences of text reduction. *Across Lang. Cult.* 14, 15–46. doi: 10.1556/Acr.14.2013.1.2

Vandekerckhove, R. (2009). Dialect loss and dialect vitality in flanders. *Int. J. Sociol. Lang.* 196/197, 73–97. doi: 10.1515/IJSL.2009.017

Vashistha, A., Sethi, P., and Anderson, R. (2017). "Respeak: a voice-based, crowd-powered speech transcription system," in *Proceedings of the 2017. CHI Conference on Human Factors in Computing Systems* (Denver, CO), 1855–1866.

Willemyns, R. (2003). "Dutch," in *Germanic standardization: Past to present*, eds A. Deumert and W. Vandenbussche (Amsterdam: John Benjamins), 93–125.

Yilmaz, E., McLaren, M., van den Heuvel, H., and van Leeuwen, D. (2018). Semi-supervised acoustic model training for speech with code-switching. *Speech Commun.* 105, 12–22. doi: 10.1016/j.2018.10.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/frai.2020.00039

A New Acoustic-Based Pronunciation Distance Measure

Martijn Bartelds[1]*, *Caitlin Richter*[2], *Mark Liberman*[2] *and Martijn Wieling*[1]

[1] *Center for Language and Cognition, Faculty of Arts, University of Groningen, Groningen, Netherlands,* [2] *Department of Linguistics, University of Pennsylvania, Philadelphia, PA, United States*

We present an acoustic distance measure for comparing pronunciations, and apply the measure to assess foreign accent strength in American-English by comparing speech of non-native American-English speakers to a collection of native American-English speakers. An acoustic-only measure is valuable as it does not require the time-consuming and error-prone process of phonetically transcribing speech samples which is necessary for current edit distance-based approaches. We minimize speaker variability in the data set by employing speaker-based cepstral mean and variance normalization, and compute word-based acoustic distances using the dynamic time warping algorithm. Our results indicate a strong correlation of $r = -0.71$ ($p < 0.0001$) between the acoustic distances and human judgments of native-likeness provided by more than 1,100 native American-English raters. Therefore, the convenient acoustic measure performs only slightly lower than the state-of-the-art transcription-based performance of $r = -0.77$. We also report the results of several small experiments which show that the acoustic measure is not only sensitive to segmental differences, but also to intonational differences and durational differences. However, it is not immune to unwanted differences caused by using a different recording device.

Keywords: acoustic measure, acoustic features, foreign accent, mel-frequency cepstral coefficients, pronunciation, spoken language processing, validation

OPEN ACCESS

Edited by:
James N. Stanford,
Dartmouth College, United States

Reviewed by:
Jason A. Shaw,
Yale University, United States
Joshua Waxman,
Yeshiva University, United States

****Correspondence:***
Martijn Bartelds
m.bartelds@rug.nl

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *29 December 2019*
Accepted: *04 May 2020*
Published: *29 May 2020*

Citation:
Bartelds M, Richter C, Liberman M and Wieling M (2020) A New Acoustic-Based Pronunciation Distance Measure. Front. Artif. Intell. 3:39. doi: 10.3389/frai.2020.00039

INTRODUCTION

The strength of foreign accent in a second language is mainly caused by the first language background of non-native speakers, and is influenced by a wide variety of variables with the most valuable predictor being the age of second-language learning (Asher and García, 1969; Leather, 1983; Flege, 1988; Arslan and Hansen, 1997). Understanding the factors that affect the degree of foreign accent may be essential for second language teaching, and knowledge about the acoustic features of foreign-accented speech can improve speech recognition models (Arslan and Hansen, 1996; Piske et al., 2001). Computational methods that investigate foreign accent strength are, however, scarce.

Studies that investigate and compare different pronunciations often use transcribed speech (Nerbonne and Heeringa, 1997; Livescu and Glass, 2000; Gooskens and Heeringa, 2004; Heeringa, 2004; Wieling et al., 2011; Chen et al., 2016; Jeszenszky et al., 2017). For example, Kessler (1995) presented the Levenshtein distance for finding linguistic distances between language varieties. To calculate the Levenshtein distance, speech samples have to be manually transcribed using a phonetic alphabet, but this process is very time consuming and labor intensive (Hakkani-Tür et al., 2002; Novotney and Callison-Burch, 2010). Furthermore, transcribing speech is prone to errors, and interference from transcriber variation might lead to a sub-optimal distance calculation when differences in transcribers' habits cannot be distinguished from differences in speakers' productions (Bucholtz, 2007). Another limitation of this approach is that the set of discrete symbols used in phonetic transcriptions is unable to capture all the acoustic details that are relevant for studying accented pronunciations (Cucchiarini, 1996). As Mermelstein (1976) notes, transcribing speech results in a loss of information whereby perceptually distinct differences between sounds diminish or largely disappear. For example, problems may arise when fine-grained pronunciation differences cannot be represented by the set of transcription symbols (Duckworth et al., 1990), or when an important dimension of difference between accents is their use of tone, but no tone or pitch information is transcribed (Heeringa et al., 2009). It is therefore potentially useful to develop an acoustic-only method to study pronunciation differences, such as foreign accent strength in the speech of non-native speakers. Fine-grained characteristics of human speech are preserved in the speech representations, while at the same time a time consuming and costly process may be omitted.

To evaluate computational methods of determining accent differences, validation against reliable data regarding these differences is necessary, which usually consists of comparing the automatically obtained ratings to human judgments of accent strength. Derwing and Munro (2009) stress the importance of including human judgments, since these provide the most appropriate method to evaluate these measurement techniques. Studies that compare human perceptual judgments to a computational difference measure which is not based on the alignment of phonetic transcriptions are uncommon, despite the potential advantages of this approach. This may be due to the challenges of directly comparing speech samples, as there exists a considerable amount of variability in the signal. A substantial amount of variability in the structure of a speech signal is also dependent on non-linguistic characteristics of the speakers, which may mask relevant phonetic information in acoustic measurements (Goslin et al., 2012). For example, Heeringa et al. (2009) calculated speaker-dependent pronunciation distances for a set of fifteen speakers from different Norwegian varieties and for a subset of 11 female speakers. The Manhattan distance was computed between the frequency values of the first three formants per vowel in each word. Correlations between their procedure and human judgments of native-likeness only ranged from $r = 0.36$ to $r = 0.60$ ($p < 0.001$). Given that they only obtained a moderate correlation with the human judgments, their acoustic-based measure could not serve as a reliable alternative to transcription-based methods for assessing accent differences.

The primary goal of this study is therefore to develop an improved acoustic pronunciation distance measure that computes pronunciation distances without requiring phonetic transcriptions. To assess whether the acoustic distance measure is a valid measurement technique to measure accent strength (compared to native speakers), we compare the acoustic distances to a collection of human native-likeness judgments that were collected by Wieling et al. (2014) to evaluate a phonetic transcription-based method. The core of the acoustic distance measure is to use dynamic time warping (DTW) to compare non-native accented American-English to native-accented American-English speech samples represented as Mel-frequency cepstral coefficients (MFCCs). In short, our approach consists of obtaining word-level acoustic differences, which are averaged to obtain speaker-based acoustic differences. To make the comparison less dependent on individual speaker characteristics, we use speaker-based cepstral mean and variance normalization before calculating the word-level acoustic differences. We evaluate the method by comparing the acoustic distances to both transcription-based pronunciation distances and human perception. To better understand what (desired and less desired) differences are captured by our acoustic difference measure, we conduct several small-scale experiments.

MATERIALS AND METHODS

Speech Accent Archive

We use data from the Speech Accent Archive, which contains over 2000 speech samples from both native and non-native American-English speakers (Weinberger, 2015). For each participant an acoustic voice recording of the same standard 69-word-paragraph is present. The paragraph is primarily composed of common English words, and contains a wide variety of consonants and vowels that can be found in the English language. The paragraph is shown in (1).

(1) *Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.*

The availability of data from both native and non-native speakers of American-English enables us to compare the accents of a broad range of different speakers of English (Weinberger and Kunath, 2011). Speech samples from 280 non-native American-English speakers make up our target non-native speaker data set, and 115 speech samples from U.S.-born L1 speakers of English serve as our reference native speaker data set. For each non-native speaker the goal is to determine how different that speaker's pronunciation is on average from the native American-English speakers in the reference native speaker data set. We do not rely on choosing a single native American-English reference speaker, as there is considerable regional variability in the data set. The native American-English speakers who rated the non-native speech samples also had different regional backgrounds.

The data we include in this study is similar to the data used for evaluating a transcription-based measurement in the study of Wieling et al. (2014). As in some cases a word was produced twice by a speaker, or two words were merged into one word, we removed duplicate words from the speech samples by deleting one of the repeated words, and merged words were split such that each speech sample consisted of 69 separate words.

Our data set contains slightly more male speakers (206) than female speakers (189). The average age of all speakers in our data set is 32.6 years with a standard deviation of 13.5 years. In the target non-native speaker data set, the average age of starting to learn English is 10.5 years with a standard deviation of 6.6 years. The 280 non-native English speakers have a total of 99 different native languages. The most frequent native languages in the target data set of non-native English speakers are Spanish ($N = 17$), French ($N = 13$), and Arabic ($N = 12$). A total of 46 languages is only spoken by a single speaker.

Human Judgments of Native-Likeness

Perceptual data have been widely used to assess the degree of foreign-accentedness (Koster and Koet, 1993; Munro, 1995; Magen, 1998; Munro and Derwing, 2001). We therefore use human judgments of native-likeness that were collected in the study of Wieling et al. (2014). They created an online questionnaire in which native speakers of American-English were asked to rate the accent strength of 50 speech samples extracted from the Speech Accent Archive. The degree of native-likeness of the speech samples was judged on a 7-point Likert scale. A score of 1 was assigned to a speaker that was perceived as very foreign-sounding, and a score of 7 was assigned to a speaker that was perceived as having native American-English speaking abilities. The speech samples presented to the participants were not duplicated, so each participant rated each sample at most once. The set of samples available for different participants to judge was changed several times during the period the questionnaire was online. To increase the reliability of the ratings, not all speech samples from the Speech Accent Archive were included in the questionnaire, so that each speech sample could be judged by multiple participants. It was also not compulsory to rate all 50 samples, because the participants could decide to rate a subset of the speech samples.

The questionnaire of Wieling et al. (2014) was distributed by asking colleagues and friends to forward it to native speakers of American-English. The questionnaire was also mentioned in a blog post of Mark Liberman[1] which led to a considerable amount of responses. In total, 1,143 participants provided native-likeness ratings (57.6% men and 42.4% woman). On average, they rated 41 samples with a standard deviation of 14 samples. The participants had a mean age of 36.2 years with a standard deviation of 13.9 years, and people most frequently came from California (13.2%), New York (10.1%), and Massachusetts (5.9%).

Experimental Setup

Segmentation

We obtain acoustic distances comparing speakers from the target data set to the speakers in the reference data set. The data sets we use contain recordings of the entire 69 word paragraph (henceforth referred to as the complete speech sample). These complete speech samples do not only contain the 69 word pronunciations, but also speech disfluencies. Examples of these disfluencies include, but are not limited to, (filled) pauses, false starts, word order changes, or mispronunciations.

To only compare corresponding segments of speech, we segment each complete speech sample into words. While this segmentation procedure may be performed manually, this is very time consuming (Goldman, 2011). We therefore employ the Penn Phonetics Lab Forced Aligner (P2FA) to time-align the speech samples with a word-level orthographic transcription (Yuan and Liberman, 2008). The P2FA is an automatic phonetic alignment toolkit that is based on the Hidden Markov Toolkit (HTK). Prior to creating the forced alignments, we resample each of the speech samples to 11,025 Hz (Yuan and Liberman, 2008). The forced alignment approach identifies the word boundaries in the speech samples, and by using this information we automatically divide the complete speech samples of the target and reference data set into separate words. Each word corresponds to a word from the elicitation paragraph presented in (1). In this way, we also remove non-speech elements that exist between these word boundaries, preventing them from entering the acoustic distance calculation. After the forced alignment procedure, we have a target data set that for each of the 280 speakers contains 69 segmented speech samples, as well as a reference data set of 115 speakers with for each speaker 69 corresponding segmented speech samples. A detailed explanation of the theoretical framework behind the forced alignment procedure is provided in the studies of Young and Young (1993) and Bailey (2016).

Feature Representation

For each segmented speech sample in both data sets, we calculate a numerical feature representation based on Mel-frequency cepstral coefficients (MFCCs). MFCCs have shown their robustness, as these speech features are widely used as representations of phonetic content in automatic speech recognition systems (Davis and Mermelstein, 1980).

We visualize the computation of each MFCC feature representation in **Figure 1**. The first, commonly executed, step in calculating this numerical feature representation is to compensate for the negative spectral slope of each speech sample (Sluijter and Van Heuven, 1996). The nature of the glottal pulses causes voiced segments in the audio signal to contain more energy at the lower frequencies compared to the higher frequencies (Vergin and O'Shaughnessy, 1995). We remove some of these glottal effects from the spectrum of the vocal tract by applying a filter to the audio signal (see Equation 1). This filter emphasizes the higher frequencies, and as a result a more balanced spectrum of the speech sample is obtained. This is usually referred to as the pre-emphasis step (Muda et al., 2010).

$$H(z) = 1 \text{ - } 0.97 * z^{\text{-}1} \tag{1}$$

We then divide each speech sample into short frames of time using a windowing function. These frames of analysis are important since the characteristics of an audio signal are fairly stable when a short frame of time is taken into account (Zhu

[1] https://languagelog.ldc.upenn.edu/nll/?p=3967, May 19, 2012, "Rating American English Accents."

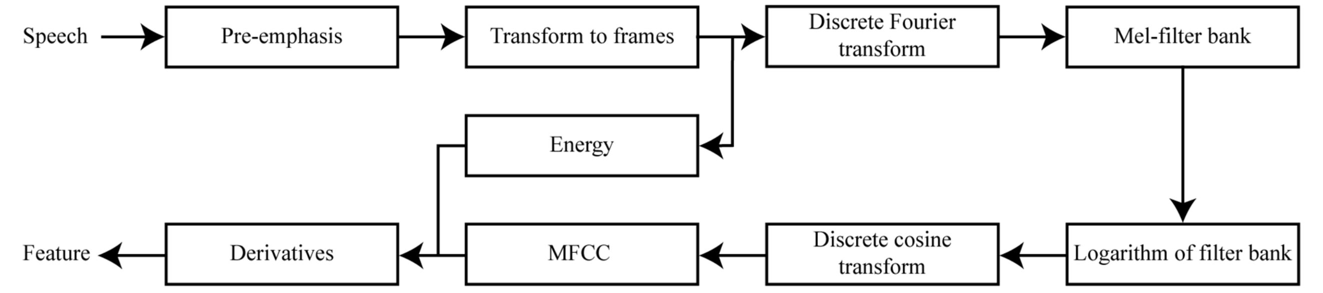


FIGURE 1 | Diagram visualizing the features used in our acoustic distance algorithm.

and Alwan, 2000). We create overlapping frames of a 25 ms time interval using a 10 ms step size. A set of cepstral coefficients is computed for each of these windowed frames per speech sample. The Hamming windowing function is used to extract each frame from the audio signal (Deller et al., 1993).

The Discrete Fourier Transform (DFT) is then taken from each of these windowed frames to transform the audio signal from the time domain to the frequency domain (Zheng et al., 2001). Taking the DFT of the windowed frames is related to the way sound is perceived by human beings. The oscillation of the human cochlea depends on the frequency of incoming sounds, and these oscillations inform the human brain that certain frequencies are present in the audio signal. With the application of DFT, the process that occurs within the human auditory system is simulated (Dave, 2013).

After the DFT is taken from the windowed frames, the Mel spectrum is computed. The DFT-transformed audio signal is modified by passing it through a collection of triangular band-pass filters. These filters are also known as the Mel filter bank, and each processes frequencies that occur within a certain range while discarding frequencies that are outside that range (Muda et al., 2010). The Mel filter bank then provides information about the amount of energy that is present near certain frequency regions (Rao and Manjunath, 2017). The width of the filter banks is determined via Mel-scaling. Units on the Mel scale are based on the way frequencies are perceived by the human auditory system. These Mel units do not correspond to tone frequencies in a linear way, as the human auditory system does not perceive frequencies linearly. Instead, the Mel scale is composed such that the frequencies below 1,000 Hz are approximately linearly spaced, and the frequencies above 1,000 Hz are distributed according to a logarithmic scale (Stevens et al., 1937).

The first filters of the Mel-filter bank are most strict, since the low frequencies are the most informative in speech perception (Raut and Shah, 2015). The energy of voiced speech is mostly concentrated at the lower frequencies (Seltzer et al., 2004). After the DFT-transformed audio signal goes through the triangular-shaped band-pass filters, the logarithm is taken of the energies that are returned by the Mel-filter bank. This procedure is also in accordance with the human auditory system, since humans do not perceive the loudness of an incoming audio signal linearly. The final result of this procedure is a signal that is represented in the cepstral domain (Oppenheim and Schafer, 2004).

The logarithmically transformed filter bank energy representations do, however, overlap. To provide a solution to the overlapping filter banks, the discrete cosine transform (DCT) is computed from the logarithmically transformed filter bank output. The result of the DCT is a set of cepstral coefficients. Following an established standard, we chose to solely include the first 12 cepstral coefficients and energy in each frame, which characterize the most relevant information of the speech signal (Picone, 1993). In addition, we calculate the first-order and second-order derivatives from each of the cepstral coefficients and energy features (Furui, 1981). We therefore have 12 first-order and 12 second-order derivatives that are associated with the 12 cepstral coefficients, and one first-order and second-order derivative related to the energy feature. These first-order and second-order derivatives, or (double) delta coefficients, model the changes between the frames over time (Muda et al., 2010). A total of 39 coefficients is computed at each 10 ms step per speech sample, to represent the most important phonetic information embedded within each 25 ms windowed frame. The MFCC feature representation per segmented speech sample is obtained by concatenating its corresponding vectors of 39 coefficients computed for each of the windowed frames.

Normalization

Ganapathy et al. (2011) and Shafik et al. (2009) showed that the quality of the MFCC feature representation is highly influenced by the presence of noise in the speech samples. To reduce the effect of noise, cepstral mean and variance normalization is applied to the feature representations (Auckenthaler et al., 2000). In addition to the robustness in the presence of noisy input, cepstral mean and variance normalization reduces the word error rate in automatic speech recognition implementations, and improves the generalization across speakers (Haeb-Umbach, 1999; Molau et al., 2003; Tsakalidis and Byrne, 2005). Adank et al. (2004) showed that cepstral mean and variance normalization can be used to highlight the linguistic content of the feature representations.

We implement cepstral mean and variance normalization by applying a linear transformation to the coefficients of the MFCC feature representations (Lu et al., 2009). The MFCC feature representations are standardized per speaker by removing the speaker's mean, and scaling to unit variance. The equation that we use to calculate the cepstral mean and variance normalized feature representations is shown in Equation (2).

$$\hat{c}(i,t) = \frac{c(i,t) - \bar{c}(i,t)}{\sigma(i)} \tag{2}$$

In this equation, the i-th cepstral coefficient at time index t is represented by $c(i,t)$. The mean value of each feature representation, and the corresponding standard deviation are given by $\bar{c}(i,t)$ and $\sigma(i)$, respectively. In Equations (3) and (4), we show how the mean value and standard deviation are obtained. In these equations, N corresponds to the number of windows used in processing the speech sample.

$$\bar{c}(i,t) = \frac{1}{N} * \sum_{t=1}^{N} c(i,t) \tag{3}$$

$$\sigma(i) = \sqrt{\frac{1}{N} * \sum_{t=1}^{N} (c(i,t) - \bar{c}(i,t))^2} \tag{4}$$

Dynamic Time Warping

The acoustic word distances are computed using the dynamic time warping (DTW) algorithm. This algorithm compares two MFCC feature representations, and returns their degree of similarity as a distance score (Galbally and Galbally, 2015). DTW has already been widely used in the domain of speech recognition, and is also used for sequence comparison in many other research domains, such as computer vision and protein structure matching (Sakoe et al., 1990; Bahlmann and Burkhardt, 2004; Efrat et al., 2007).

To compare a target pronunciation with a reference pronunciation, the DTW algorithm uses the corresponding target and reference MFCC feature representations. These are shown in Equations (5) and (6).

$$\text{target} = (x_1, x_2, ..., x_n) \tag{5}$$

$$\text{reference} = (y_1, y_2, ..., y_m) \tag{6}$$

An $m * n$ cost matrix is created to align the target MFCC feature representation with the reference MFCC feature representation (Muda et al., 2010). This cost matrix is filled with the Euclidean distances between every pair of points (frames) in both the target and reference MFCC feature representations (Danielsson, 1980). For example, element (i,j) of the cost matrix contains the distance d that is given by Equation (7).

$$d(\text{target}_i, \text{reference}_j) = (\text{target}_i - \text{reference}_j)^2 \tag{7}$$

The optimal alignment between the MFCC feature representations corresponds to the shortest path through the cost matrix, and is therefore to some extent comparable to the edit distance. The DTW algorithm computes the shortest path using an iterative method that calculates the minimum cumulative distance $\gamma(i,j)$ (Keogh and Pazzani, 2001). The cumulative distance is composed of the distance in the current cell $d(\text{target}_i, \text{reference}_j)$ and the minimum of the cumulative distance found in the adjacent cells (shown in Equation 8).

$$\begin{aligned}\gamma(i,j) &= d(\text{target}_i, \text{reference}_j) \\ &+ min\big(\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\big)\end{aligned} \tag{8}$$

After the cumulative distance is computed, it is divided by the length of the target feature representation and the reference feature representation $(n + m)$. It is important to normalize the computed distances, since the speech samples we work with do not necessarily have the same length. Without normalization applied to DTW, longer alignment paths (from longer recordings) would have higher distances than shorter alignments, because they have more frames to accumulate cost (Giorgino et al., 2009).

The final speaker pronunciation distances are obtained by first calculating the acoustic distance for each of the 69 words pronounced by a non-native speaker of American-English and a single native speaker of American-English in the reference data set. We subsequently average these word-based distances to measure the between-speaker acoustic distance. The difference between the pronunciation of a non-native speaker and native American-English in general, is determined by calculating the between-speaker acoustic distances compared to all 115 native American-English speakers, and subsequently averaging these. We compute these acoustic distances for all foreign-accented speech samples by applying this same procedure to each of the 280 non-native speakers of American-English in the target data set. To evaluate our measure, the correlation between the native-likeness ratings and the acoustic distances is computed. We evaluate the impact of the (size of the) set of reference speakers, by calculating the correlation for successively smaller subsets of reference speakers.

Understanding the Acoustic Distance Measure

In addition to the main experiment, we perform a variety of other analyses to obtain a more complete understanding of the acoustic details captured by the acoustic distance measure.

First we use a multiple linear regression model to predict the human native-likeness ratings on the basis of our acoustic distance measure, but also using the transcription-based distances reported by Wieling et al. (2014), and the (manually counted) number of mispronunciations a speaker made, as these might be important for native-likeness ratings (Flege, 1981), but are not included in either of the two other measures.

Second, to assess whether our acoustic distance measure adequately captures fine-grained segmental differences, we compute acoustic differences between 10 repetitions of hVd words (e.g., [hɪd]) pronounced by a single speaker. We subsequently correlate these differences with differences based on the first and second formant measured at the mid-point of the vowel of the recordings. We follow Wieling et al. (2012) in Bark-scaling the formant-based distances. We use a total of 12 Dutch monophthongs in the vowel context (a, ɑ, ɛ, e, ø, ɪ, i, ɔ, u, o, ʏ, y). We visualize the differences (both the formant-based distances, and the acoustic-based distances) using multidimensional scaling (Torgerson, 1952).

Third and finally, to assess whether non-segmental variability is also captured by our acoustic method, we compute acoustic distances between four series of recordings (10 repetitions) of the word "living". The first and second series consisted of a normal

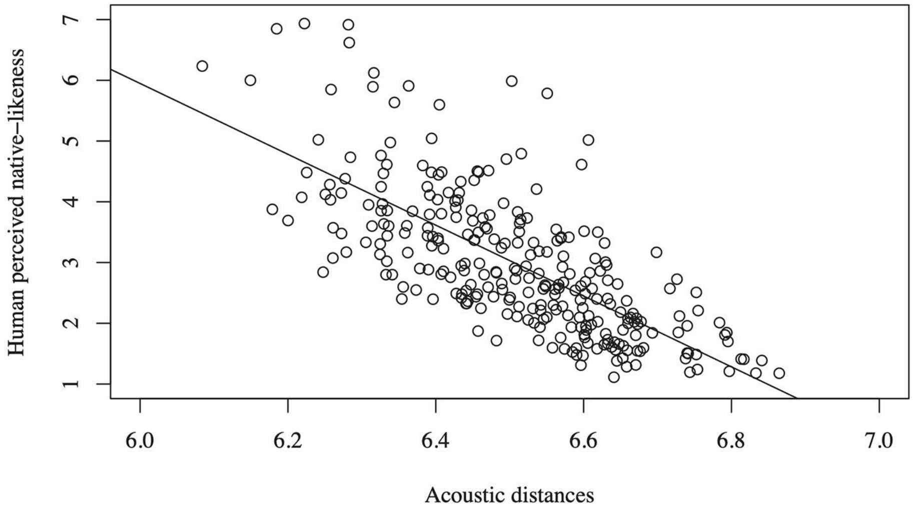


FIGURE 2 | Native-likeness ratings as a function of the computed acoustic distances (r = −0.71).

TABLE 1 | Pearson correlation coefficients r between the acoustic distances and human judgments of native-likeness depending on the size of the reference data set.

Amount of reference speakers	*r*
10	−0.68
25	−0.71
50	−0.70
75	−0.72

All correlations are significant at the $p < 0.0001$ *level.*

TABLE 2 | Pearson correlation coefficients r of acoustic distances compared to human judgments of native-likeness, using different methods to compute the acoustic distances.

Model	*r*
Baseline 1 (only segmentation)	−0.27
Baseline 2 (only normalization)	−0.63
Acoustic measure (segmentation and normalization)	−0.71

All correlations are significant at the $p < 0.0001$ *level.*

pronunciation ("living"), but recorded with two recording devices (the built-in microphone of a laptop, and the built-in microphone of a smartphone), the third series consisted of a pronunciation in which the intonation was changed ("living?"), and the fourth series consisted of a pronunciation in which the relative duration of the syllables was changed ("li_ving").

RESULTS

The correlation between the native-likeness ratings and the acoustic distances computed using our acoustic method is $r = -0.71$ ($p < 0.0001$), and therefore accounts for about half of the variance in the native-likeness ratings ($r^2 = 0.50$). **Figure 2** visualizes this correlation in a scatter plot. The acoustic distance measure tends to underestimate the native-likeness (overestimate distances) when the speech samples are rated as being very native-like.

Compared to the transcription-based method of Wieling et al. (2014), who used the Levenshtein distance incorporating automatically determined linguistically-sensible segment distances, and reported a correlation of $r = -0.77$, the performance of our measure is significantly lower (using the modified z-statistic of Steiger (1980): $z = 2.10$, $p < 0.05$).

Impact of Reference Speakers

As the set of reference speakers might affect the correlation, we evaluated the impact of reducing the set of reference speakers. The results are shown in **Table 1** and show that the correlation remains comparable, irrespective of the (size of the) reference set (i.e., $-0.68 \leq r \leq -0.72$). To assess whether language variation within the set of reference speakers might be important, we computed the acoustic distances using as our reference set ($N = 14$) only the native American-English speakers who originated from the western half of the U.S. and the English-speaking part of Canada. These areas are characterized by less dialect variation compared to the eastern half of the U.S. (Boberg, 2010). Again, this did not substantially affect the correlation, as it remained similar ($r = -0.70$).

Impact of Segmentation and Normalization

Two simplified (baseline) measures, each missing a single component of our acoustic measure, were created to assess how segmentation and cepstral mean and variance normalization of the speech samples contribute to acoustic distances that are more similar to human judgments of native-likeness. The results of this experiment is shown in **Table 2**. It is clear that not using the normalization approach is much more detrimental than not segmenting, but that the best results are obtained when doing both. The modified z-statistic of Steiger (1980) indicates that our acoustic method significantly outperforms either of the two simpler methods ($z = 4.11$, $p < 0.0001$).

Understanding the Acoustic Distance Measure

We fitted a multiple linear regression model to determine whether the acoustic distance measure and the transcription-based distance measure captured distinctive aspects of pronunciation. We also assessed the influence of the number of mispronunciations. The coefficients and associated statistics of the predictors used are shown in **Table 3**. The results show that the transcription-based distances and acoustic distances both contribute significantly to the model fit ($p < 0.05$). This is not the case for the amount of mispronunciations per speaker in the target data set ($p > 0.05$). The presented model accounts for 65% of the variation in the human judgments of native-likeness ($r^2 = 0.65$). Only using the transcription-based distance measure accounted for 60% of the variation. Consequently, our acoustic measure also seems to capture information which is not present in phonetic transcriptions.

TABLE 3 | Coefficients of a multiple regression model predicting human judgments of native-likeness.

	Estimate	Std. Error	t-value	p-value
Intercept	24.19	2.68	9.04	< 0.001
Transcription-based distances	−379.30	34.26	−11.07	< 0.001
Acoustic-based distances	−2.79	0.44	−6.35	< 0.001
Amount of mispronunciations	0.01	0.03	0.26	0.795

TABLE 4 | Averaged acoustic distances and standard errors of four variants of the word "living".

	Compared to normal pronunciation
Normal pronunciation	4.35 (0.50)
Normal pronunciation (different recording device)	6.94 (0.15)
Rising intonation	7.12 (0.13)
Lengthened first syllable	6.65 (0.13)

The results in **Table 4**, show that our acoustic measure can capture both intonation and timing differences as these lead to larger distances than comparing individual repetitions of the same word pronounced by the same speaker. However, it also shows that when recording the normal pronunciation by two microphones simultaneously, the acoustic distances between the two simultaneous recordings are higher than zero, whereas the pronunciation is in fact identical. Note that the acoustic distance when comparing the 10 normal pronunciations is also not zero, due to small deviations in the pronunciations.

Another indication of how well our acoustic measure captures segmental information is shown by the significant positive correlation of $r = 0.68$ ($p < 0.0001$) between the formant-based acoustic vowel differences and the computed acoustic differences between the hVd-words. **Figure 3** shows these relative vowel distances by using a multidimensional scaling visualization of the formant-based vowel differences (visualizing all variation) and the DTW-based vowel differences (visualizing 47% of the variation in the original differences).

DISCUSSION

We have created an acoustic-only approach for calculating pronunciation distances between utterances of the same word by different speakers. We have evaluated the measure by calculating how different the speech of non-native speakers of American-English is from native American-English speakers, and by comparing our computed results to human judgments of native-likeness. While our method is somewhat outperformed ($r = -0.71$ vs. $r = -0.77$) by the transcription-based

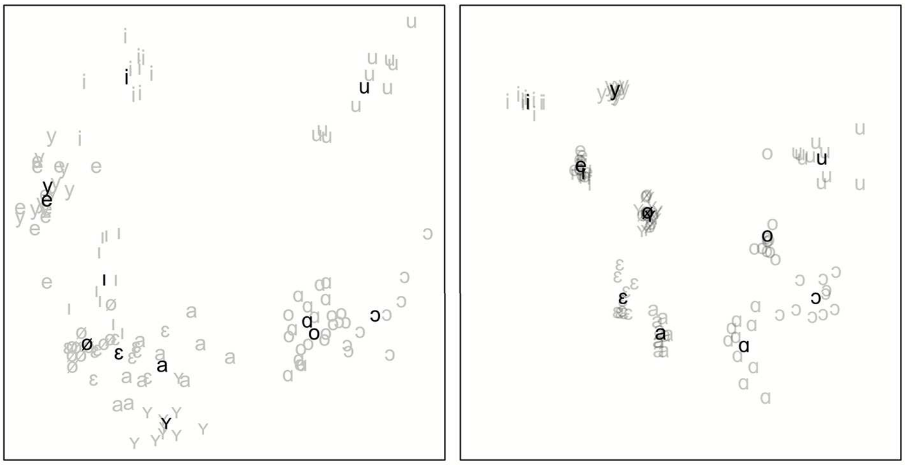


FIGURE 3 | MDS plots visualizing the acoustic vowel distances **(left)** and the formant-based vowel distances **(right)**. Individual pronunciations are shown in light gray, whereas the averages per vowel are shown in black.

method introduced by Wieling et al. (2014), our measure does not require phonetic transcriptions, whose production is time consuming and prone to errors. Given that our method is fully automatic, the trade-off in performance may be worthwhile.

Word segmentation and especially speaker-based cepstral mean and variance normalization of the MFCC speech representations were important in creating an adequate acoustic-based distance measure. These results show the importance of pre-processing continuous speech samples, as the comparison of pronunciations in speech samples is most reliable when it is based on comparable and normalized segments of speech that we obtain from word-level forced alignment.

The multiple regression model showed that the acoustic distance measure explained variance not accounted for by the transcription-based distance measure. Particularly, our further experiments showed that our measure is both sensitive to timing and intonation differences. However, the measure is also sensitive to different recording devices, which is undesirable and may partly explain why the method is outperformed by the transcription-based method. While the MFCC feature representation with cepstral mean and variance normalization attempts to minimize non-linguistic confounds, it is only partly successful, as a computational representation of general phonetic information remains a difficult issue in speech processing technology (Gemmeke et al., 2011).

Consequently, future work should investigate whether other acoustic (pre-processing) techniques may improve our acoustic measure. For example, contextual acoustic encoding techniques related to word embeddings like *wav2vec* and *vq-wav2vec* may highlight acoustic details that are linguistically relevant (Baevski et al., 2019; Schneider et al., 2019). Additionally, generating a shared phonetic space through which two speech samples may be compared (Ryant and Liberman, 2016) may be useful. Nevertheless, our work serves as a useful and promising starting point for a fully automatic acoustic pronunciation distance measure.

DATA AVAILABILITY STATEMENT

The code and datasets generated for this study are available at: https://github.com/Bartelds/acoustic-distance-measure.

AUTHOR CONTRIBUTIONS

MB and MW conceptualized the research. MB and CR designed and conducted the experiments, performed data analysis, and data visualization. MB wrote the first version of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The research discussed in this paper has been funded by the Center for Groningen Language and Culture, Faculty of Arts, University of Groningen. CR was supported by NSF grant DGE-1321851.

ACKNOWLEDGMENTS

The authors would like to express their gratitude toward Hedwig Sekeres for her contribution visualizing the formant-based vowel distances and assessing the amount the mispronunciations in the data set.

REFERENCES

Adank, P., Smits, R., and Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *J. Acoust. Soc. Am.* 116, 3099–3107. doi: 10.1121/1.1795335

Arslan, L. M., and Hansen, J. H. (1996). Language accent classification in American english. *Speech Commun.* 18, 353–367. doi: 10.1016/0167-6393(96)00024-6

Arslan, L. M., and Hansen, J. H. (1997). A study of temporal features and frequency characteristics in American english foreign accent. *J. Acoust. Soc. Am.* 102, 28–40. doi: 10.1121/1.419608

Asher, J. J., and García, R. (1969). The optimal age to learn a foreign language. *Modern Lang. J.* 53, 334–341. doi: 10.1111/j.1540-4781.1969.tb04603.x

Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Process.* 10, 42–54. doi: 10.1006/dspr.1999.0360

Baevski, A., Schneider, S., and Auli, M. (2019). vq-wav2vec: self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453.*

Bahlmann, C., and Burkhardt, H. (2004). The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 299–310. doi: 10.1109/TPAMI.2004.1262308

Bailey, G. (2016). "Automatic detection of sociolinguistic variation using forced alignment," in *University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44)* (York), 10–20.

Boberg, C. (2010). *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511781056

Bucholtz, M. (2007). Variation in transcription. *Discourse Stud.* 9, 784–808. doi: 10.1177/1461445607082580

Chen, N. F., Wee, D., Tong, R., Ma, B., and Li, H. (2016). Large-scale characterization of non-native mandarin Chinese spoken by speakers of European origin: analysis on icall. *Speech Commun.* 84, 46–56. doi: 10.1016/j.specom.2016.07.005

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clin. Linguist. Phonet.* 10, 131–155. doi: 10.3109/02699209608985167

Danielsson, P.-E. (1980). Euclidean distance mapping. *Comput. Graph. Image Process.* 14, 227–248. doi: 10.1016/0146-664X(80)90054-4

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *Int. J. Adv. Res. Eng. Technol.* 1, 1–4.

Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* 28, 357–366. doi: 10.1109/TASSP.1980.1163420

Deller, J. R. Jr., Proakis, J. G., and Hansen, J. H. (1993). *Discrete Time Processing of Speech Signals*. Upper Saddle River, NJ: Prentice Hall PTR.

Derwing, T. M., and Munro, M. J. (2009). Putting accent in its place: rethinking obstacles to communication. *Lang. Teach.* 42, 476–490. doi: 10.1017/S026144480800551X

Duckworth, M., Allen, G., Hardcastle, W., and Ball, M. (1990). Extensions to the international phonetic alphabet for the transcription of atypical speech. *Clin. Linguist. Phonet.* 4, 273–280. doi: 10.3109/02699209008985489

Efrat, A., Fan, Q., and Venkatasubramanian, S. (2007). Curve matching, time warping, and light fields: new algorithms for computing similarity between curves. *J. Math. Imaging Vis.* 27, 203–216. doi: 10.1007/s10851-006-0647-0

Flege, J. E. (1981). The phonological basis of foreign accent: a hypothesis. *Tesol Quart*. 15, 443–455. doi: 10.2307/3586485

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in english sentences. *J. Acoust. Soc. Am*. 84, 70–79. doi: 10.1121/1.396876

Furui, S. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Acoust. Speech Signal Process*. 29, 342–350. doi: 10.1109/TASSP.1981.1163605

Galbally, J., and Galbally, D. (2015). A pattern recognition approach based on DTW for automatic transient identification in nuclear power plants. *Ann. Nucl. Energy* 81, 287–300. doi: 10.1016/j.anucene.2015.03.003

Ganapathy, S., Pelecanos, J., and Omar, M. K. (2011). "Feature normalization for speaker verification in room reverberation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Prague), 4836–4839. doi: 10.1109/ICASSP.2011.5947438

Gemmeke, J. F., Virtanen, T., and Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio Speech Lang. Process*. 19, 2067–2080. doi: 10.1109/TASL.2011.2112350

Giorgino, T. et al. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw*. 31, 1–24. doi: 10.18637/jss.v031.i07

Goldman, J.-P. (2011). "Easyalign: an automatic phonetic alignment tool under praat," in *Proceedings of the Annual Conference of the International Speech Communication Association* (Florence: INTERSPEECH), 3233–3236.

Gooskens, C., and Heeringa, W. (2004). Perceptive evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Lang. Variat. Change* 16, 189–207. doi: 10.1017/S0954394504163023

Goslin, J., Duffy, H., and Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain Lang*. 122, 92–102. doi: 10.1016/j.bandl.2012.04.017

Haeb-Umbach, R. (1999). "Investigations on inter-speaker variability in the feature space," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*(Phoenix, AZ: IEEE), 397–400. doi: 10.1109/ICASSP.1999.758146

Hakkani-Tür, D., Riccardi, G., and Gorin, A. (2002). "Active learning for automatic speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 4* (Orlando, FL), doi: 10.1109/ICASSP.2002.5745510

Heeringa, W., Johnson, K., and Gooskens, C. (2009). Measuring norwegian dialect distances using acoustic features. *Speech Commun*. 51, 167–183. doi: 10.1016/j.specom.2008.07.006

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (Ph.D. thesis). Citeseer.

Jeszenszky, P., Stoeckle, P., Glaser, E., and Weibel, R. (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in swiss german. *J. Linguist. Geogr*. 5, 86–108. doi: 10.1017/jlg.2017.5

Keogh, E. J., and Pazzani, M. J. (2001). "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining* (Philadelphia, PA), 1–11. doi: 10.1137/1.9781611972719.1

Kessler, B. (1995). "Computational dialectology in Irish gaelic," in *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics* (Dublin: Morgan Kaufmann Publishers Inc.), 60–66. doi: 10.3115/976973.976983

Koster, C. J., and Koet, T. (1993). The evaluation of accent in the english of Dutchmen. *Lang. Learn*. 43, 69–92. doi: 10.1111/j.1467-1770.1993.tb00173.x

Leather, J. (1983). Second-language pronunciation learning and teaching. *Lang. Teach*. 16, 198–219. doi: 10.1017/S0261444800010120

Livescu, K., and Glass, J. (2000). "Lexical modeling of non-native speech for automatic speech recognition," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), Vol. 3* (Istanbul), 1683–1686. doi: 10.1109/ICASSP.2000.862074

Lu, X., Matsuda, S., Unoki, M., Shimizu, T., and Nakamura, S. (2009). "Temporal contrast normalization and edge-preserved smoothing on temporal modulation structure for robust speech recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (Taipei), 4573–4576. doi: 10.1109/ICASSP.2009.4960648

Magen, H. S. (1998). The perception of foreign-accented speech. *J. Phonet*. 26, 381–400. doi: 10.1006/jpho.1998.0081

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recogn. Artif. Intell*. 116, 374–388.

Molau, S., Hilger, F., and Ney, H. (2003). "Feature space normalization in adverse acoustic conditions," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03)* (Hong Kong). doi: 10.1109/ICASSP.2003.1198866

Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: ratings of filtered speech. *Stud. Second Lang. Acquis*. 17, 17–34. doi: 10.1017/S0272263100013735

Munro, M. J., and Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of l2 speech the role of speaking rate. *Stud. Second Lang. Acquis*. 23, 451–468. doi: 10.1017/S0272263101004016

Nerbonne, J., and Heeringa, W. (1997). "Measuring dialect distance phonetically," in *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonolby* (Stroudsburg, PA: Association for Computational Linguistics (ACL)), 11–18.

Novotney, S., and Callison-Burch, C. (2010). "Cheap, fast and good enough: automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Los Angeles, CA: Association for Computational Linguistics), 207–215.

Oppenheim, A. V., and Schafer, R. W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Process. Mag*. 21, 95–106. doi: 10.1109/MSP.2004.1328092

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proc. IEEE* 81, 1215–1247. doi: 10.1109/5.237532

Piske, T., MacKay, I. R., and Flege, J. E. (2001). Factors affecting degree of foreign accent in an l2: a review. *J. Phonet*. 29, 191–215. doi: 10.1006/jpho.2001.0134

Rao, K. S., and Manjunath, K. (2017). *Speech Recognition Using Articulatory and Excitation Source Features*. Cham: Springer. doi: 10.1007/978-3-319-49220-9

Raut, S. P., and Shah, D. S. N. (2015). Voice biometric system for speaker authentication. *Int. J. Comput. Appl*. 975:8887.

Ryant, N., and Liberman, M. (2016). "Large-scale analysis of spanish/s/-lenition using audiobooks," in *Proceedings of Meetings on Acoustics 22ICA, Vol. 28* (Buenos Aires: ASA), 060005. doi: 10.1121/2.0000500

Sakoe, H., Chiba, S., Waibel, A., and Lee, K. (1990). Dynamic programming algorithm optimization for spoken word recognition. *Read. Speech Recogn*. 159:224. doi: 10.1016/B978-0-08-051584-7.50016-4

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*. doi: 10.21437/Interspeech.2019-1873

Seltzer, M. L., Raj, B., and Stern, R. M. (2004). A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Commun*. 43, 379–393. doi: 10.1016/j.specom.2004.03.006

Shafik, A., Elhalafawy, S. M., Diab, S., Sallam, B. M., and El-Samie, F. A. (2009). A wavelet based approach for speaker identification from degraded speech. *Int. J. Commun. Netw. Inform. Secur*. 1:52.

Sluijter, A. M., and Van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am*. 100, 2471–2485. doi: 10.1121/1.417955

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull*. 87:245. doi: 10.1037/0033-2909.87.2.245

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am*. 8, 185–190. doi: 10.1121/1.1915893

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika* 17, 401–419. doi: 10.1007/BF02288916

Tsakalidis, S., and Byrne, W. (2005). "Acoustic training from heterogeneous data sources: experiments in mandarin conversational telephone speech transcription," in *Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, Vol. 1* (Philadelphia, PA). doi: 10.1109/ICASSP.2005.1415150

Vergin, R., and O'Shaughnessy, D. (1995). "Pre-emphasis and speech recognition," in *Proceedings 1995 Canadian Conference on Electrical* *and Computer Engineering, Vol. 2* (Montreal, QC), 1062–1065. doi: 10.1109/CCECE.1995.526613

Weinberger, S. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

Weinberger, S. H., and Kunath, S. A. (2011). The speech accent archive: towards a typology of english accents. *Lang. Comput. Stud. Pract. Linguist*. 73:265. doi: 10.1163/9789401206884_014

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., and Nerbonne, J. (2014). Measuring foreign accent strength in english: validating levenshtein distance as a measure. *Lang. Dyn. Change* 4, 253–269. doi: 10.1163/22105832-004 02001

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *J. Phonet*. 40, 307–314. doi: 10.1016/j.wocn.2011.12.004

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS ONE* 6:e23613. doi: 10.1371/journal.pone.0023613

Young, S. J., and Young, S. (1993). *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. University of Cambridge; Department of Engineering Cambridge.

Yuan, J., and Liberman, M. (2008). Speaker identification on the scotus corpus. *J. Acoust. Soc. Am*. 123:3878. doi: 10.1121/1.2935783

Zheng, F., Zhang, G., and Song, Z. (2001). Comparison of different implementations of MFCC. *J. Comput. Sci. Technol*. 16, 582–589. doi: 10.1007/BF02943243

Zhu, Q., and Alwan, A. (2000). "On the use of variable frame rate analysis in speech recognition," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3* (Istanbul), 1783–1786.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/frai.2020.00038

Toward "English" Phonetics: Variability in the Pre-consonantal Voicing Effect Across English Dialects and Speakers

James Tanner[1]*, *Morgan Sonderegger*[1], *Jane Stuart-Smith*[2] and *Josef Fruehwald*[3]

[1] Department of Linguistics, McGill University, Montreal, QC, Canada, [2] Glasgow University Laboratory of Phonetics, University of Glasgow, Glasgow, United Kingdom, [3] Department of Linguistics, University of Kentucky, Kentucky, KY, United States

Recent advances in access to spoken-language corpora and development of speech processing tools have made possible the performance of "large-scale" phonetic and sociolinguistic research. This study illustrates the usefulness of such a large-scale approach—using data from multiple corpora across a range of English dialects, collected, and analyzed with the SPADE project—to examine how the pre-consonantal Voicing Effect (longer vowels before voiced than voiceless obstruents, in e.g., *bead* vs. *beat*) is realized in spontaneous speech, and varies across dialects and individual speakers. Compared with previous reports of controlled laboratory speech, the Voicing Effect was found to be substantially smaller in spontaneous speech, but still influenced by the expected range of phonetic factors. Dialects of English differed substantially from each other in the size of the Voicing Effect, whilst individual speakers varied little relative to their particular dialect. This study demonstrates the value of large-scale phonetic research as a means of developing our understanding of the structure of speech variability, and illustrates how large-scale studies, such as those carried out within SPADE, can be applied to other questions in phonetic and sociolinguistic research.

Keywords: voicing effect, English, phonetic variability, Bayesian modeling, dialectal variation, speaker variability

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Joshua Waxman,
Yeshiva University, United States
George Bailey,
University of York, United Kingdom

****Correspondence:***
James Tanner
james.tanner@mail.mcgill.ca

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *24 December 2019*
Accepted: *01 May 2010*
Published: *29 May 2020*

Citation:
Tanner J, Sonderegger M, Stuart-Smith J and Fruehwald J (2020) Toward "English" Phonetics: Variability in the Pre-consonantal Voicing Effect Across English Dialects and Speakers. Front. Artif. Intell. 3:38. doi: 10.3389/frai.2020.00038

1. INTRODUCTION

There exist a large number of well-studied properties of speech that are known to vary across languages and communities of speakers, which have long been of interest to sociolinguists and phoneticians. One dimension of this variability, which is the focus of this study, is that of variation *within languages*: across dialects and their speakers. For example, the deletion of word-final /t/ and /d/ segments (in e.g., *mist*, *missed*) has been shown to vary across a wide range of dialects and speech communities (e.g., Labov et al., 1968; Guy, 1980; Tagliamonte and Temple, 2005), as have the dialect-specific realization of English vowels (e.g., Thomas, 2001; Clopper et al., 2005; Labov et al., 2006), and variation in the degree of aspiration in English voiced and voiceless stops (e.g., Docherty, 1992; Stuart-Smith et al., 2015; Sonderegger et al., 2017). The study of this kind of variation provides a means of understanding the sources and structures of variability within languages: both in how particular dialects may systematically differ from each other, and how the variable realization of speech sounds maps to speakers' cognitive representation of language and speech (Liberman et al., 1967; Lisker, 1985; Kleinschmidt, 2018). Despite decades of research, however, there is much we do not know about the scope, extent, and structure of this kind of language-internal variability. Within the phonetic literature, most research has focused on highly-controlled speech styles in 'laboratory settings', generally focusing on a single dialect in each study; much of the work focusing on phonetic variability in spontaneous speech is on single dialects (e.g., Ernestus et al., 2015). The sociolinguistic and dialectological literatures have often examined spontaneous speech, with some notable cross-dialectal studies (e.g., Clopper et al., 2005; Labov et al., 2006; Jacewicz and Fox, 2013), but nonetheless primarily focus on variation in vowel quality. Increasingly, however, research within phonetics and sociophonetics is being performed at a larger scale *across* speech communities (Labov et al., 2006, 2013; Yuan et al., 2006, 2007; Yuan and Liberman, 2014; Coleman et al., 2016; Liberman, 2018), driven by the development of new speech processing tools and data sharing agreements. This "large-scale" approach is applied here to one such well-studied variable, the pre-consonantal voicing effect, as a means of characterizing its degree and structure of variability in a single phonetic effect across English dialects and speakers.

The pre-consonantal voicing effect (henceforth *Voicing Effect*, VE) refers to vowels preceding voiced obstruents being consistently longer than their voiceless counterparts, such as the differences in *beat-bead* and *mace-maze* (House and Fairbanks, 1953; House, 1961). The VE has been reported—to greater or lesser extent—in a range of languages (Zimmerman and Sapon, 1958; Chen, 1970), though varies in size based on properties of the phonetic environment, such as whether the obstruent is a stop or fricative, the height of the vowel, and many others (Klatt, 1973; Crystal and House, 1982; Port and Dalby, 1982). The evidence for the English VE to date is sourced predominantly from laboratory studies of highly-controlled speech, often in citation form, recorded from small numbers of often standard General American English speakers (e.g., Rositzke, 1939; House and Fairbanks, 1953; Peterson and Lehiste, 1960; House, 1961; Crystal and House, 1982; Luce and Charles-Luce, 1985). On the basis of this evidence, the VE has been noted for being particularly large in English relative to other languages (Zimmerman and Sapon, 1958; Chen, 1970), and has long been suggested as a prominent cue to consonant voicing in English (Denes, 1955; Klatt, 1973). This in turn has motivated claims that the VE is learned in English, as opposed to being a low-level phonetic property in other languages (Fromkin, 1977; Keating, 2006; Solé, 2007). At the same time, numerous questions about the nature and extent of the VE in English remain unexplored. In this study, we will examine the variability in the VE across a range of English dialects, focusing on the following two research questions: (1) *how large is the VE as realized in spontaneous English speech?* and (2) *how much does the VE vary across dialects and speakers?* In addressing these questions, we hope to gain insight into a number of open issues, including the extent to which there is a single "English" VE or whether dialects differ in the magnitude of the effect, as well as the range of VE sizes across individual speakers of a given dialect.

This paper answers these questions by taking a "large-scale" approach to the study of the VE. Concretely, this refers to the use of a large amount of acoustic data, collected from a large number of speakers across a range of English dialects. This analysis falls within the framework of the *SPeech Across Dialects of English* (SPADE) project (Sonderegger et al., 2019, https://spade.glasgow.ac.uk/), which aims to consider phonetic and phonological variation in British and North American English across time and space through the use of automated acoustic analysis of features across English dialects occurring in many corpora. The methodological and research goals of the SPADE project are exemplified through this study of the English VE, specifically by the use of multiple corpora of diverse sources and structures, and the use of linguistic and acoustic analysis via the *Integrated Speech Corpus ANalysis* (ISCAN) tool (McAuliffe et al., 2019), developed as part of the broader SPADE project. Both the volume and complexity of the resulting data and the goals of the study motivate the need for appropriately-flexible approaches to the statistical analysis: specifically, the data is statistically analyzed using Bayesian regression models (Carpenter et al., 2017), which enable us to accurately estimate the size of the VE across dialects and speakers directly, whilst controlling for the complex nature of the spontaneous speech data.

The structure of this paper is as follows. Section 2 outlines previous work on the VE, and some of the outstanding questions related to our current understanding of its variability. Section 3 describes the data: the corpora of different dialects from SPADE. Sections 4, 5 describe the methodological approach: the process of acoustic and statistical analysis of the data. The results of this analysis are reported in section 6, and then discussed with respect to our specific research questions in section 7 and concluding in section 8.

2. THE VOICING EFFECT (VE)

The observation that vowels preceding voiced obstruents are consistently longer than before voiceless obstruents was first noted in early phonetics textbooks (e.g., Sweet, 1880; Kenyon, 1940; Thomas, 1947; Jones, 1948) and in preliminary experimental work from the first half of the twentieth century (Heffner, 1937; Rositzke, 1939; Hibbitt, 1948). Studies explicitly manipulating the VE in English observed an effect of around 1.45—that is, vowels before voiced consonants were longer than before voiceless consonants by a ratio of around 2:3 (House and Fairbanks, 1953; House, 1961), and this effect was a cue to the voicing of the obstruent (Denes, 1955; Lisker, 1957; Raphael, 1972).

In these studies, VE was shown to be affected by consonant manner: namely, that fricatives showed a smaller or minimal VE compared to stops (Peterson and Lehiste, 1960), and less-robustly cued the voicing of the final consonant (Raphael, 1972). Initial studies of connected speech suggested that the size of the VE in this type of speech is more variable: VEs in carrier sentences are similar to those in isolated words (Luce and Charles-Luce, 1985)[1] whilst vowels in read or spontaneous speech exhibit smaller VE sizes of around 1.2, and a negligible VE for fricatives (Crystal and House, 1982; Tauberer and Evanini, 2009). VE size

[1] Harris and Umeda (1974), in their study of overall vowel duration, attribute this difference to a "mechanical" prosody as a consequence of numerous repetitions.

is also modulated by the overall length of the vowel, which is hypothesized to be due to an intrinsic incompressibility of the vowel, limited by the minimal time required to perform the articulatory motor commands necessary for vowel production (Klatt, 1976). This general suggestion has been supported by observations that VE is smaller for unstressed and phrase-medial vowels (Umeda, 1975; Klatt, 1976), and vowels produced at a faster speech rate (Crystal and House, 1982; Cuartero, 2002). The VE is thus modulated by a range of phonetic factors, and largely predict a reduction of VE size in instances where vowels are generally shorter; vowels that undergo "temporal compression" have a reduced capacity to maintain a large VE size, and so VE is minimized. As these effects have only been investigated in laboratory speech, it is not clear whether the size and direction of these effects are maintained in less-controlled spontaneous speech styles.

Examining the VE across languages, Zimmerman and Sapon (1958) first observed that whilst English speakers produced a robust VE, Spanish speakers did not modulate vowel length in the same way, though this study did not control for the syllabic structure of test items. Comparing across English, French, Russian, and Korean, Chen (1970) observed that all four languages produced a VE size of at least 1.1, though all languages had different VE sizes (English = 1.63, French = 1.15, Russian = 1.22, Korean = 1.31). This was interpreted as evidence that VE is a phonetically-driven effect with additional language-specific phonological specification (Fromkin, 1977). Mack (1982), comparing English and French monolinguals with bilinguals, observed that English monolinguals maintained a substantially larger VE than French monolinguals, whilst the French-English bilinguals also produced the shorter French-style pattern instead of adapting to the larger English VE pattern. Keating (1985) suggested that VE is "phonetically-preferred," though ultimately controlled by the grammar of the particular language. English, then, is expected to have a larger VE than other languages, though it is not known if the English VE is of a comparable size in spontaneous speech.

The work discussed above has not differentiated between varieties of English, and cross-linguistic comparisons of VE have presumed that a single "English" VE size exists. Little work has focused on variation in VE across English dialects beyond a small number of studies on specific dialects. One dialect group of interest has been Scottish Englishes and the application of the Scottish Vowel Length Rule (SVLR), where vowels preceding voiced fricatives and morpheme boundaries are lengthened, whilst all other contexts have short vowels (Aitken, 1981), and hence do not show the VE. In studies of the SVLR, some East Coast Scotland speakers show some evidence of the VE in production (Hewlett et al., 1999), whilst VE-like patterns were not observed in spontaneous Glaswegian (Rathcke and Stuart-Smith, 2016). On the other hand, studies of African American English (AAE) have claimed that voiced stops undergo categorical devoicing in this variety, which has resulted in additional vowel lengthing before voiced stops to maintain the pre-consonantal voicing contrast (Holt et al., 2016; Farrington, 2018). Only one study has previously compared the VE *across* English dialects in spontaneous speech. Tauberer and Evanini (2009), using interview data from the *Atlas of North American English* (Labov et al., 2006), observe that North American English dialects vary in their VE values, ranging from 1.02 to 1.33, and that dialects with shorter vowels on average (New York City) also show a smaller-than-average VE size (1.13). Moreover, despite recognition that individual speakers may exhibit variability in their VE sizes (Rositzke, 1939; Summers, 1987), no study has formally examined the extent of variability across speakers, nor how dialects may differ in the degree of VE variability amongst its speakers. The two patterns observed for Scottish and African American English suggest that English dialects can maintain relatively "small" (or no), and "large" VEs, respectively; we know little about the degree of VE variability beyond these dialects without a controlled study across multiple English varieties, which is one of the goals of this study.

Whilst a large number of studies on the VE have provided useful information for its realization in English and other languages, there are still a range of outstanding questions that can be addressed through a large-scale cross-dialectal approach. To what extent is the VE a *learned* property of a given language, compared with an *automatic* consequence of low-level phonetic structure? Much of the discussion with respect to variation in VE has revolved around differences across *languages* (Chen, 1970; Keating, 1985), which may differ both in their phonetic realization of segments but also the phonological representation of those segments. In this sense, examining VE variability internal to a language (i.e., across *dialects*) potentially avoids this problem; the specification of phonological categories—here, the voicing status of final obstruents—are expected be largely consistent within a language, meaning that language-internal variability may be driven by only differences in phonetic implementation.

Little is known about how English dialects may vary in their implementation of the VE, and so a range of possibilities exist for how dialects might compare. One possibility is that, with the exception of varieties with specific phonological rules interacting with the VE, dialects might cluster around a single "English" VE value, potentially of the size reported in the previous literature. Such a finding would support the previous approach in the literature, in terms of English compared to other languages, and suggest that dialects do not differ in how the final voicing contrast is phonetically implemented. Alternatively, dialects may differ gradiently from each other, and so may show a continuum of possible dialect-specific VE sizes. If dialects do differ in their VE size in this way, this would suggest that the previous literature on the VE in "English" accounts for just a fraction of the possible VE realizations across English, and would provide evidence that individual English dialects differ in their phonetic implementation of an otherwise "phonological" contrast (Keating, 1984, 1985).

Similarly, little is known about how individual speakers vary in the VE, and what the overall distribution of speaker VE sizes is. Synchronic variability across speakers is one of the key inputs to sound change (Ohala, 1989; Baker et al., 2011), and also defines the limits of a speech community, i.e., speakers who share sociolinguistic norms in terms of production and social evaluation (e.g., Labov, 1972). Whilst dialects may differ in the realization of segments or the application of phonological processes, dialect-internal variability is potentially more limited if a phonetic alternation such as the VE is critical to speech community membership.

3. DATA FOR THIS STUDY

The varieties of English included in this study are from North America, Great Britain, and Ireland. For the purposes of this study, North American dialects refer to the regions of the United States and Canada outlined in *The Atlas of North American English*, which is based around phonetic, not lexical, differences between geographic regions (Labov et al., 2006; Boberg, 2018). For Canadian data specifically, the primary distinction was made between "urban" and "rural" speakers, based on its relative importance noted in comparison to much weaker geographic distinctions, at least for the corpus which makes up most Canadian data in this study (Rosen and Skriver, 2015). Within the British and Irish groups, dialects from England in this study are defined in terms of Trudgill's dialectal groupings (Trudgill, 1999), which groups regions in terms of both phonological and lexical similarity. Due to the lack of geographical metadata for speakers from Ireland and Wales, these dialects were simply coded as "Ireland" and "Wales" directly. Scottish Englishes are grouped based on information from *The Scottish National Dictionary*[2]. The data used in this study comes from the SPADE project, which aims to bring together and analyze over 40 speech corpora covering English speech across North America, the United Kingdom, and Ireland. In this study, we analyze data from 15 of these corpora, which together cover 30 different English dialects from these regions, comprised of speech from interviews, conversations, and reading passages. A basic description of each of these corpora is given below, outlining the type of speech and phonetic alignment tools used.

- *Audio British National Corpus* (AudioBNC, Coleman et al., 2012): The spoken sections of the British National Corpus, originally containing speech from over 1,000 speakers. However, due to a range of recording issues (e.g., overlapping speech, background noise, microphone interference), a large portion of the corpus is inaccurately aligned. In order to define a subset of the AudioBNC which maximizes the accuracy of the alignment, utterances were kept if they met a number of criteria: the utterance length was greater than one second, that the utterance contained at least two words, that the mean harmonics-to-noise ratio of the recording was at least 5.6, and that the mean difference in segmental boundaries between the alignment and a re-alignment with the Montreal Forced Aligner (MFA, McAuliffe et al., 2017a) was at most 30 ms[3]. 50 TextGrids from the remaining data were manually checked and deemed to be as approximately accurate as that of normal forced-alignment.
- *Brains in Dialogue* (Solanki, 2017): recordings of 24 female Glaswegian speakers producing spontaneous speech in a laboratory setting. There are 12 recordings for each speaker, which were aligned with LaBB-CAT (Fromont and Hay, 2012).
- *Buckeye* (Pitt et al., 2007): spontaneous interview speech of 40 speakers from Columbus Ohio, recorded in 1990s–2000s. The Buckeye corpus is hand-corrected with phonetic transcription labels: these were converted back to phonological transcriptions in order to be comparable with data from the other corpora.
- *Corpus of Regional African American Language* (CORAAL, Kendall and Farrington, 2018): spontaneous sociolinguistic interviews with 100 AAE speakers from Washington DC, Rochester NY, and Princeville NC, recorded between 1968 and 2016, and aligned with the MFA.
- *Doubletalk* (Geng et al., 2013): recordings of paired speakers carrying out a variety of tasks in order to elicit a range of styles/registers in a discourse/interactive situation. Ten speakers make up five pairs where one member is a speaker of Southern Standard British English and the other member is a speaker of Scottish English.
- *Hastings* (Holmes-Elliott, 2015): recordings of sociolinguistic interviews with 46 speakers from Hastings in the south east of England, male and female, aged from 8 to 90, aligned using FAVE (Rosenfelder et al., 2014).
- *International Corpus of English—Canada* (ICE-Canada, Greenbaum and Nelson, 1996): interview and broadcast speech of Canadian English, recorded in the 1990s across Canada, and aligned using the MFA. Speaker dialect was defined in terms of their city or town of origin. In this study, we coded a speaker as "urban" if their birthplace was a large Canadian city.
- *Canadian Prairies* (Rosen and Skriver, 2015): Spontaneous sociolinguistic interviews, recorded between 2010 and 2016, with speakers of varying ethnic backgrounds from the provinces of Alberta and Manitoba, conducted as part of the Language in the Prairies project, and was aligned using the MFA.
- *Modern RP* (Fabricius, 2000): reading passages by Cambridge University students recorded in 1990s and 2000s. The speakers were chosen for having upper middle-class backgrounds as defined by at least one parent having a professional occupation along with the speaker also having attended private schooling. The data used in this study come from a reading passage aligned with FAVE.
- *Philadelphia Neighborhood Corpus* (PNC, Labov and Rosenfelder, 2011): sociolinguistic interviews with 419 speakers from Philadelphia, recorded between 1973 and 2013, and were aligned with FAVE.
- *Raleigh* (Dodsworth and Kohn, 2012): semi-structured sociolinguistic interviews of 59 White English speakers in Raleigh, North Carolina, born between 1955 and 1989, and aligned with the MFA.
- *Santa Barbara* (Bois et al., 2000): spontaneous US English speech, recorded in the 1990s and 2000s, from a range of speakers of different regions, genders, ages, and social backgrounds.

[2] Part of *The Dictionary of the Scots Language* (https://dsl.ac.uk/).

[3] We are grateful to Michael Goodale for designing and performing this filtering protocol.

- *The Scottish Corpus of Texts and Speech* (SCOTS, Anderson et al., 2007): approximately 1,300 written and spoken texts (23% spoken), ranging from informal conversations, interviews, etc. Most spoken texts were recorded since 2000.
- *Sounds of the City* (SOTC, Stuart-Smith et al., 2017): vernacular and standard Glaswegian from 142 speakers over 4 decades (1970s–2000s), collected from historical archives and sociolinguistic surveys, aligned using LaBB-CAT.
- *Switchboard* (Godfrey et al., 1992): 2,400 spontaneous telephone conversations between random participants from the multiple dialect regions in the United States on a variety of topics, containing data from around 500 speakers.

The goals of this study are to examine the size and variability in the English VE in spontaneous speech, and in variation in the VE across dialects and individual speakers. Specifically, the kind of dialectal variability being addressed in this study is that of *regional* variability: variability by race or ethnicity is not being directly considered in this study, with the exception of three African American English varieties, given the particular observations about AAE with respect to the VE (Holt et al., 2016; Farrington, 2018). This study also does not focus on differences according to age, either age-grading or apparent/real-time change in the VE over time; only speech data recorded since 1990s was included; the other data recorded prior to 1990 was excluded from further analysis. Analysis of the role of age and time in the VE in these English dialects remains a subject for future study.

4. DATA ANALYSIS

Having collected and organized the speech data into dialects, it is then possible to extract and acoustically analyze the data in the study: that is, going from raw data (audio and transcription files) to datasets which can be statistically analyzed. As the corpora differ in their formats—the phone labels used, organization of speaker data, etc.—modifying the acoustic analysis procedure for each different corpus format would be both labor and time-intensive, as well as increase the risk that the analysis itself differed across corpora. In order to standardize the acoustic analysis across corpora, the *Integrated Speech Corpus Analysis* (ISCAN) tool was developed for use in this kind of cross-dialectal study in the context of the SPADE project. This section provides a brief overview of the ISCAN system: see McAuliffe et al. (2017b, 2019) and the ISCAN documentation page for details of the implementation[4].

The process of deriving a dataset from raw corpus files consists of three major steps. In the first step, individual speech corpora (in the form of sets of audio-transcription pairs) are *imported* into a graph database format, where each transcription file is minimally composed of word and phone boundaries (e.g., word-level and phone-level tiers in a TextGrid), and these word-phone relationships are structurally-defined in the database (i.e., that each phone belongs to a word). Importers have been developed for a range of standard automatic aligners, including all formats of corpora described in section 3. Corpora, represented in database format, can then be further *enriched* with additional structure, measurements, and linguistic information. For example, utterances can be defined as groups of words (separated silence of a specified length, e.g., 150 ms), syllables can be defined as a property between groups of adjacent phones. Once the database has been enriched with utterance and syllable information, speech rate (often defined as syllables per second within an utterance) can be calculated and included in the database. Similarly, information about words (such as frequency) or speakers (such as gender, age, dialect etc.) can be added to the corpus from metadata files. Once a corpus has been sufficiently enriched with linguistic and acoustic information, it is then possible to perform a *query* on the corpus at a given level of analysis. This level of analysis refers to the level of the hierarchy on which the resulting datafile should use as the main level of observation, for example individual phones, syllables, or utterances. Filters can be applied to a query to restrict it to the particular contexts of interest, for example, including only syllables occurring at the right edge of an utterance, or vowels followed by a specific subset of phone types (e.g., obstruents). Finally, the resulting query can then be *exported* into a data format (currently CSV only) for further analysis.

Each corpus was processed using the ISCAN software pipeline, and then combined into a single "master" dataset, containing all phonetic, dialect, and speaker information from all of the analyzed corpora necessary to carry out the analysis of the VE below. As the vowel duration annotations from the corpora (except for Buckeye) were created via forced alignment with a minimum duration of 10 ms and a time-step of 30 ms, any token with a vowel duration below 50 ms was excluded from further study, as is common in acoustic studies of vowel formants to exclude heavily reduced vowels (Dodsworth, 2013; Fruehwald, 2013). To reduce the additional prosodic and stress effects on vowel duration, the study only included vowels from monosyllabic words occurring phrase-finally, where a phrase is defined as a chunk of speech separated by 150 ms of silence. Raw speech rate was calculated as syllables per second within a phrase, from which two separate speech rates were derived. First, a mean speech rate for each speaker was calculated, which reflects whether a speaker is a "fast" or "slow" speaker overall. From that mean speech rate, a local speech rate was calculated as the raw rate for the utterance subtracted from the given speaker's mean. This local speech rate can be interpreted as how fast or slow that speaker produced the vowel within that particular phrase *relative* to their average speech rate (Sonderegger et al., 2017; Cohen Priva and Gleason, 2018). Word frequency was defined using the SUBTLEX-US dataset (Brysbaert and New, 2009). The final dataset contained 229,406 vowel tokens (1,485 word types) from 1,964 speakers from 30 English dialects. **Table 1** shows the number of speakers and tokens for each dialect, and how many speakers/tokens were derived from each speech corpus.

5. STATISTICAL ANALYSIS

The research goals of this study focus on the size and variability of the VE in English spontaneous speech, and how the VE

[4] https://iscan.readthedocs.io/

TABLE 1 | Number of speakers and tokens per dialect (left), and by corpora from which each dialect was derived.

Region	Dialect	*n* Speakers	*n* tokens	Corpus	*n* speakers	*n* tokens
North America	Canada (rural)	52	9,313	Canadian Prairies	44	8,316
				ICE-Canada	8	997
	Canada (urban)	64	12,124	Canadian Prairies	56	11,939
				ICE-Canada	8	185
	Midwest US	40	5,567	Buckeye	40	5,567
	New England	24	1,336	Santa Barbara	7	174
				Switchboard	17	1,162
	North Midland US	46	3,084	Switchboard	46	3,084
	Northern Cities US	21	1,377	Santa Barbara	21	1,377
	Northern US	58	3,086	Switchboard	58	3,086
	NYC	25	1,477	Santa Barbara	6	158
				Switchboard	19	1,319
	Philadelphia	371	59,581	PNC	371	59,581
	Princeville NC (AAE)	71	6,759	CORAAL	17	6,759
	Raleigh US	92	3,282	Raleigh	92	3,282
	Rochester NY (AAE)	14	6,308	CORAAL	14	6,308
	South Midland US	108	8,188	Switchboard	108	8,188
	Southern US	44	2,738	Santa Barbara	6	345
				Switchboard	38	2,393
	Washington DC (AAE)	50	21,205	CORAAL	50	21,205
	Western US	100	5,456	Santa Barbara	50	2,900
				Switchboard	50	2,556
United Kingdom & Ireland	Central Scotland	24	2,426	SCOTS	24	2,426
	East Central England	51	2544	Audio BNC	51	2,544
	East England	229	20,727	Audio BNC	132	6,622
				Doubletalk	5	726
				Hastings	44	12,642
				ModernRP	48	737
	Edinburgh	18	1,148	SCOTS	18	1148
	Glasgow	177	33,938	Brains in Dialogue	23	9,210
				SCOTS	27	2,294
				SOTC	127	2,2434
	Insular Scotland	8	351	SCOTS	8	351
	Ireland	19	624	Audio BNC	19	624
	Lower North England	60	3,325	Audio BNC	60	3,325
	North East England	17	488	Audio BNC	17	488
	Northern Scotland & Islands	33	2280	SCOTS	33	2,280
	Scotland	70	3,468	Audio BNC	65	2,633
				Doubletalk	5	835
	South West England	50	2,067	Audio BNC	50	2,067
	Wales	41	2,524	Audio BNC	41	2,524
	West Central England	41	2,615	Audio BNC	41	2,615
Total		1,964	229,406			

varies across dialects and speakers. These goals motivate an approach of *estimating* the size of the VE in these contexts, rather than testing whether the VE "exists" or not. Whilst controlled laboratory experiments are explicitly designed to balance across these contexts (by including matching numbers of tokens with stops vs. fricatives, using words with similar frequency, etc.), spontaneous speech taken from corpora is rarely balanced in this sense: some speakers speak more than others, have different conversations leading to some combinations of segments occurring infrequently relative to others, speakers manage properties of their speech (such as speech rate) for communicative purposes which are generally absent in laboratory studies. In trying to obtain an accurate estimate of the VE (or indeed any other linguistic property), the unbalanced nature of spontaneous speech motivates the need for a statistical approach where individual factors of interest (e.g., obstruent manner of articulation, dialects, etc.) can be explored whilst controlling for the influence of other effects. This approach—the use of multiple regression to model corpus data—is now common in phonetics and sociolinguistic research (e.g., Tagliamonte and Baayen, 2012; Roettger et al., 2019), but has not, to our knowledge, been used to analyze multiple levels of variability in the VE.

In this study, this approach to estimation is performed using Bayesian regression modeling. Whilst other multifactorial statistical models would also be valid, Bayesian models provide us with some advantages that make the goal of estimating the size of the VE easier. Mixed-models are ideal for use in this study, as these capture variability at multiple levels (the VE overall, across dialects, across speakers) and this variability is of direct interest for our research questions. Bayesian mixed models resemble more traditional linear mixed-effects (LME) models approaches commonly used in linguistic and phonetic research, such as those performed with the *lme4* package (Bates et al., 2015), though differ in a few key respects. First, Bayesian models make it easy to calculate the *range* of possible VE sizes in each context, as opposed to a single value that would be output in LME models: whilst LME models provide ranges for "fixed" effects (across all dialects/speakers), Bayesian models provide a range of possible sizes for each level (i.e., an individual dialect). In a Bayesian model, all parameters (coefficients) in the model are assumed to have a *prior* distribution of possible values, reflecting which effect sizes are believed to be more or less likely, before examining the data itself. The output of a Bayesian model is a set of *posterior* distributions, which result from combining the priors and the likelihood of observing the data. Each model parameter has its own posterior distribution, which each represent the range of values for that parameter that is consistent with both the modeled data, conditioned on prior expectations about likely values, and the structure of the model itself. Bayesian models are well-suited to the task in this study, as they allow for flexible fitting of model parameters, and allow the complex random-effects structures which are often recommended for fitting statistically-conservative models (Barr et al., 2013), but which often fail to converge in LME models (Nicenboim and Vasishth, 2016). See Vasishth et al. (2018) for an introduction to Bayesian modeling applied to phonetic research.

A Bayesian mixed model of log-transformed vowel duration was fit using *brms* (Bürkner, 2018): a R-based front-end for the Stan programming language (Carpenter et al., 2017), containing the following population-level ("fixed effects") predictors: the **voicing** and **manner** of the following obstruent, vowel **height** (high vs. non-high), the lexical **class** of the word (lexical vs. functional), both **mean** and **local** speech rates, and lexical **frequency**. To observe how compression of the vowel influences VE size, interactions between all of these factors with obstruent voicing were also included. The continuous predictors (both speech rates, frequency), were centered and divided by two standard deviations (Gelman and Hill, 2007). The two-level factors (obstruent voicing, manner, vowel height, lexical class) were converted into binary (0,1) values and then centered.

The group-level ("random effects") structure of the model contained the complete set of model predictors for both dialects and speakers, nested within dialects. These terms capture two kinds of variability in the VE size: for each individual dialect, as well as the degree of variability across speakers—the nesting of speaker term inside dialects can be interpreted as capturing the variability in the size of the VE across speakers *within* a given dialect. Given the expectation that both the overall vowel duration (represented by the intercept) and the manner of the obstruent would affect the size of the VE, correlation terms between the intercept and both the consonant voicing and manner predictors, as well as for the interaction *between* the voicing and manner predictors, were included for both dialects and speakers. Random intercepts were included for words and phoneme labels, also nested within dialects. The model was fit using 8,000 samples across 4 Markov chains (2000/2000 warmup/sample split per chain) and was fit with weakly informative "regularizing" priors (Nicenboim and Vasishth, 2016; Vasishth et al., 2018): the intercept prior used a normal distribution with a mean of 0 and a standard deviation of 1 [written as $Normal(0, 1)$]; the other fixed effects parameters used $Normal(0, 0.5)$ priors, with the exception of the obstruent voicing parameter which used a $Normal(0.1, 0.2)$ prior[5]. The group-level (for dialects, speakers) parameters used the *brms* default prior of a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 10. The correlations between group-level effects used the LKJ (Lewandowski et al., 2009) with $\zeta = 2$, which gives lower prior probability to perfect ($-1/1$) correlations, as recommended by Vasishth et al. (2018).

6. RESULTS

The results in this study will be reported in the context of the two main research questions concerning VE variability (1) in spontaneous speech, and (2) across English dialects and individual speakers. The results are reported for each effect in terms of the median value with 95% credible intervals (CrIs), and the probability of that effect's direction. These values enable us to understand the *size* of the effect (i.e., the change in vowel duration) and the confidence in the effect's predicted direction. The strength of evidence for an effect is distinct from the strength of the effect itself: to value the strength of evidence for an effect, we follow the recommendations of Nicenboim and Vasishth (2016) and consider there to be *strong* evidence of an effect if the 95% credible interval does not include 0, and *weak* evidence for an effect if 0 is within the 95% CrI but the probability of the effect's direction is at least 95% (i.e., that there is <5% probability that the effect changes direction). Evaluating the strength of an effect

[5] The values chosen for the obstruent voicing parameter reflect the decision to allow a wide range of possible VE sizes, including values both above and below those reported in the previous literature. A sensitivity analysis was performed using an additional model fit with a "uniform" flat prior for the obstruent voicing parameter, which returned VE values differing by an order of 10^{-3}, suggesting that the decision for the weakly-informative prior did not adversely affect the reported results.

TABLE 2 | Posterior mean ($\hat{\beta}$), estimated error, upper & lower credible intervals, and posterior probability of the direction of each population-level parameter included in the model of log-transformed vowel duration.

Parameter	$\hat{\beta}$	Est.Error	95% CrI	Pr($\hat{\beta}$ <> 0)
Intercept	−1.99	0.02	[−2.03, −1.96]	1
Obstruent voicing	0.14	0.03	[0.09, 0.19]	1
Obstruent manner	0.05	0.02	[0.02, 0.08]	1
Vowel height	−0.22	0.02	[−0.25, −0.18]	1
Lexical class	−0.14	0.03	[−0.21, −0.08]	1
Speech rate (mean)	−0.22	0.01	[−0.24, −0.20]	1
Speech rate (local)	−0.28	0.01	[−0.30, −0.26]	1
Lexical frequency	−0.05	0.01	[−0.08, −0.03]	1
Voicing : Manner	−0.04	0.03	[−0.10, 0.02]	0.91
Voicing : Height	0.07	0.02	[0.02, 0.11]	1
Voicing : Class	−0.07	0.03	[−0.13, 0.00]	0.97
Voicing : Mean rate	−0.01	0.01	[−0.03, 0.01]	0.77
Voicing : Local rate	−0.06	0.01	[−0.08, −0.03]	1
Voicing : Frequency	−0.07	0.02	[−0.11, −0.03]	1

is determined with respect to effect sizes previously reported for laboratory (e.g., House and Fairbanks, 1953; House, 1961) and connected speech (Crystal and House, 1982; Tauberer and Evanini, 2009). The degree of variability across dialects can be compared with the findings of Tauberer and Evanini (2009); as there is no known comparison for speaker variability, this will be compared to variability across dialects as an initial benchmark.

6.1. The Voicing Effect in Spontaneous Speech

Table 2 reports the population-level ("fixed") effects for each parameter in the fitted model. The "overall" VE size averaging across dialects, which is between 1.09 and 1.2, is estimated to be smaller than reported in previous laboratory studies ($\hat{\beta} = 0.14$, CrI = [0.09, 0.19], $\Pr(\hat{\beta} > 0) = 1$)[6] and more consistent with VE sizes reported in studies of connected and spontaneous speech (Crystal and House, 1982; Tauberer and Evanini, 2009).

Looking at how the overall VE size for all dialects is modulated by phonetic context, there is weak evidence that the manner of the following obstruent modulates VE size ($\hat{\beta} = -0.04$, CrI = [−0.10, 0.02], $\Pr(\hat{\beta} < 0) = 0.91$): whilst stops appear to have a larger VE size (**Figure 1**, top left), the uncertainty in VE size for each obstruent manner (represented by the spread of the credible intervals) suggests that it is possible there is no difference in VE size between both obstruent manners. Whilst high vowels are shown to be shorter than non-high vowels overall ($\hat{\beta} = -0.22$, CrI = [−0.25, −0.18], $\Pr(\hat{\beta} < 0) = 1$), there is strong evidence that high vowels have a larger VE than non-high vowels ($\hat{\beta} = 0.07$, CrI = [0.02, 0.11], $\Pr(\hat{\beta} > 0) = 1$). There is a similarly strong effect for lexical class ($\hat{\beta} = -0.07$, CrI = [−0.13, 0.00], $\Pr(\hat{\beta} < 0) = 0.97$), where functional words have smaller VEs than open-class lexical items (**Figure 1**, top right). Lexical frequency also has a strong and evident effect on VE size ($\hat{\beta} = -0.07$, CrI = [−0.11, −0.03], $\Pr(\hat{\beta} < 0) = 1$), where higher-frequency words have smaller VEs than their lower-frequency counterparts (**Figure 1**, bottom left), whilst local speech rate also reduces VE size ($\hat{\beta} = -0.06$, CrI = [−0.08, −0.03], $\Pr(\hat{\beta} < 0) = 1$; **Figure 1**, bottom middle). For mean speaking rate, however, the effect on VE is both small with weak evidence ($\hat{\beta} = -0.01$, CrI = [−0.03, 0.01], $\Pr(\hat{\beta} < 0) = 0.77$): this is reflected in **Figure 1** (bottom right), where the difference between faster and slower speakers has a negligible effect on VE size. These results generally suggest that shorter vowels (within-speaker) tend to have smaller VE sizes, consistent with the temporal compression account (Klatt, 1973): the apparent exception to this is the relationship between VE size and vowel height, which is addressed in section 7.

6.2. Voicing Effect Across Dialects and Speakers

Turning to dialectal variability in VE, we observe that the dialect variation in VE (the dialect-level standard deviation, $\hat{\sigma}_{dialect}$) is between 0.07 and 0.12: this can be interpreted as meaning that the difference in VE size between a "low" and "high" VE dialect is between 32 and 61%[7] (**Table 3**). This is comparable with the range of possible values for the overall VE (between 0.09 and 0.19, **Table 2**, row 2). To understand whether this constitutes a "large" degree of variability, one metric is to assess whether a "low VE" dialect would actually have a reversed effect direction (voiceless > voiced), which is tested by subtracting $2 \times \hat{\sigma}_{dialect}$ from the overall VE size and comparing to 0. There is little evidence that dialects differ enough to change direction ($\hat{\beta} = -0.05$, CrI = [−0.09, 0], $\Pr(\hat{\beta} > 0) = 0.06$), which suggests that whilst individual dialects differ in the *size* of the VE, no dialect fully differs in the *direction* of the effect (i.e., no dialect's credible interval is fully negative).

Another way of understanding the degree of dialectal variability in VE is to examine the predicted VE for individual dialects. As shown in **Figure 2**, dialects appear to differ gradiently from each other, ranging from dialects with effectively-null VE to those with strong evidence for large VEs. The Scottish dialects of Central Scotland and Edinburgh have VEs of at most 1.06 and 1.09, respectively, based on their upper credible interval value, whilst their median values (indicated by the points in **Figure 2**) indicate that the most likely VE size is around 0 (Central Scotland: $\hat{\beta} = 0.99$, CrI = [0.93, 1.06]; Edinburgh: $\hat{\beta} = 1.01$, CrI = [0.93, 1.09]): indeed, all Scottish dialects have a predicted VE size of 1.16 at the highest, with most of these having median values <1.1 (**Table 4**). North American dialects, in contrast, all have robustly positive VE values (no credible interval crosses the 0 line) and are generally larger than the British and Irish variants, shown by the position of red (North American) and blue (United Kingdom and Ireland) points respectively in **Figure 2**. In particular, the AAE dialects have the largest VEs in the sample, which are all robustly larger than the average "English" VE size (Rochester NY: $\hat{\beta} = 1.35$, CrI = [1.27, 1.44]; Princeville NC: $\hat{\beta} =$

[6] As vowel duration was log-transformed prior to fitting, effects are interpreted by taking the exponent of the model parameter's value, e.g., $e^{0.19}$=1.2, which refers to a vowel duration increase of 20%.

[7] The value is multiplied by 4 to get the 95% range of values = $2\hat{\sigma}_{dialect}$ for both sides of the distribution = 0.28, which is then back-transformed from log via the exponential function = $e^{0.28} = 1.32$.

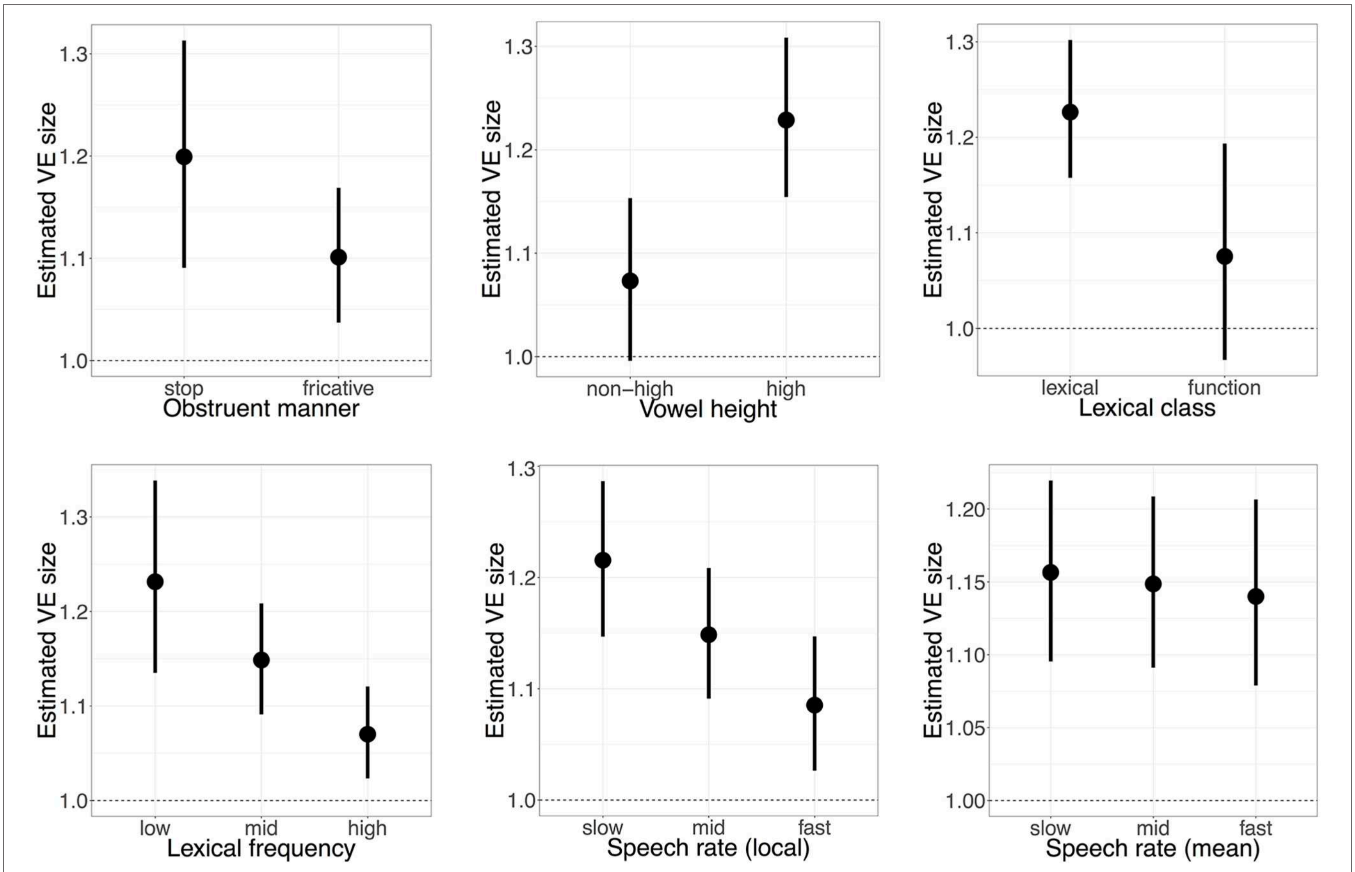


FIGURE 1 | Modulation of VE size in different phonetic contexts: obstruent manner **(Top Left)**, vowel height **(Top Middle)**, lexical class **(Top Right)**, frequency **(Bottom Left)**, local **(Bottom Middle)**, and mean **(Bottom Right)** speech rates. Points and error bars indicate the posterior mean value with 95% credible intervals, whilst holding all other predictors at their average values. Dashed line indicates no difference between vowels preceding voiced or voiceless consonants. For continuous predictors (frequency, speech rates), the estimate VE size is shown at three values for clarity.

TABLE 3 | Posterior mean ($\hat{\sigma}$), estimated error, and 95% credible intervals for dialect and speaker-level parameters related to obstruent voicing included in the model of log-transformed vowel duration.

Level	Parameter	$\hat{\sigma}$	Est.Error	95% CrI
Dialect	Intercept	0.05	0.01	[0.03, 0.07]
	Obstruent Voicing	0.09	0.01	[0.07, 0.12]
	Voicing : Manner	0.12	0.02	[0.09, 0.16]
	Voicing : Height	0.04	0.01	[0.01, 0.06]
	Voicing : Class	0.06	0.01	[0.04, 0.09]
	Voicing : Mean Rate	0.02	0.01	[0.00, 0.05]
	Voicing : Local Rate	0.05	0.01	[0.03, 0.07]
Speaker	Intercept	0.10	0.00	[0.09, 0.10]
	Obstruent Voicing	0.08	0.00	[0.07, 0.08]
	Voicing : Height	0.11	0.01	[0.10, 0.12]
	Voicing : Manner	0.11	0.01	[0.10 0.13]
	Voicing : Class	0.13	0.01	[0.11, 0.14]
	Voicing : Local Rate	0.09	0.01	[0.08, 0.11]

1.39, CrI = [1.31, 1.48]; Washington DC: $\hat{\beta} = 1.49$, CrI = [1.42, 1.56]): this is consistent with previous studies of studies on AAE, which posit that final devoicing of word-final voiced obstruents results in compensatory vowel lengthening (Holt et al., 2016; Farrington, 2018).

Turning to variability in VE across individual speakers, we observe that speakers are estimated to vary within-dialect by between 0.07 and 0.08 ($\hat{\sigma}_{speaker} = 0.08$, CrI = [0.07, 0.08]), meaning that speakers differ in their VE ratios by between 32 and 37% (**Table 3**). To put this value in context and get an impression of the size of variability across speakers, this value is compared with the degree of variability across dialects. **Figure 3** illustrates how likely the model deems different degrees of by-speaker and by-dialect variability: highest probability (darker shading) lies where by-dialect variability is greater than by-speaker variability. By the metric of between-dialect variability, **Figure 3** illustrates that whilst dialects differ in VE size, individual speakers vary little from their dialect-specific baseline value.

7. DISCUSSION

The findings from this study will be discussed with respect to the two research questions: (1) how the VE is realized in spontaneous speech, and (2) how the VE varies across dialects and speakers. The VE in English is often considered to be substantially larger

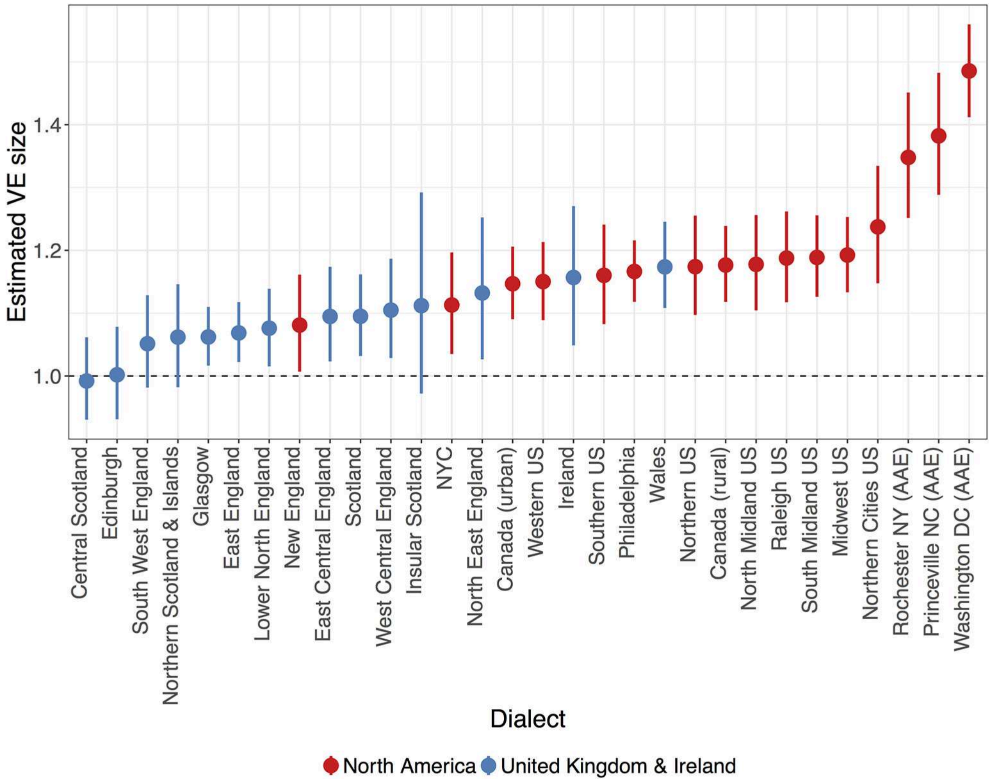


FIGURE 2 | Estimated VE size for each dialect analyzed in this study (red = North American, blue = United Kingdom and Ireland). Points and errorbars indicate the posterior mean value with 95% credible intervals, whilst holding all other predictors at their average values. Dashed line indicates no difference between vowels preceding voiced or voiceless consonants.

than in other languages (Chen, 1970) and claimed to play a significant perceptual role in cueing consonant voicing (Denes, 1955). Taken together, these observations have formed the basis for claims that the VE in English is phonologically specified beyond an otherwise phonetically-consistent acoustic property across languages (Fromkin, 1977; Keating, 1985). Previous work has focused on controlled laboratory speech, leaving open the question of how the VE is realized in spontaneous English speech.

In this study, the overall VE in spontaneous speech was observed to have a maximum size of around 1.2—substantially smaller than the 1.5 commonly reported in laboratory studies (e.g., House and Fairbanks, 1953; Peterson and Lehiste, 1960; House, 1961; Chen, 1970), and more consistent with previous research on VE in connected speech (Crystal and House, 1982; Tauberer and Evanini, 2009). Spontaneous VE size was also shown to be affected by a range of phonetic factors, such as consonant manner, vowel height, frequency, and speech rate, though the evidence for each of these effects varies substantially (section 6.1). What the effects of these phonetic factors suggest is that contexts where vowels are often shorter also have shorter VE sizes, supporting the argument of "temporal compression": that vowels which have already undergone shortening cannot be subsequently shortened further (Harris and Umeda, 1974; Klatt, 1976). An interesting exception to this finding is that the VE size was found to be larger for high vowels than non-high vowels in this study (**Figure 1**)—the direction of this effect may be counter to that predicted by temporal compression, and opens a question as to whether this and other predictions of temporal compression are straightforwardly replicable in spontaneous speech environments. The overall smaller-size and impact of phonetic factors of the VE in spontaneous speech indicates a possible fragility of the VE in spontaneous speech, in apparent contrast to the supposed perceptual importance of the VE as a cue to consonant voicing (Denes, 1955; Lisker, 1957; Raphael, 1972). This apparent conflict between the perceptual importance of the VE and its subtlety in production provides an interesting area for future work.

The fact that VE size in English differs so widely between laboratory and connected speech not only demonstrates the importance of speech style and context on phonetic realization (Labov, 1972; Lindblom, 1990), but also raises the question of "how big" the VE in English really is, or could be. If larger overall VE size is only observable in laboratory speech, it would be interesting to empirically re-evaluate the question of whether English VE is in fact larger than in other languages. For languages that exhibit smaller VEs than English in laboratory speech (Chen, 1970), it is not clear how such languages may realize the VE in more naturalistic speech. One possibility is that the VE across languages is comparatively small in spontaneous speech and similarly affected by phonetic factors; alternatively, the VE in

TABLE 4 | Estimated VE sizes (mean, estimated error, and upper and lower credible intervals) for each dialect used in this study.

Dialect	$\hat{\beta}$	Est.Error	95% CrI
Central Scotland	0.99	0.03	[0.93, 1.06]
Edinburgh	1.01	0.04	[0.93, 1.09]
South West England	1.05	0.03	[0.99, 1.12]
Glasgow	1.06	0.02	[1.02, 1.11]
Northern Scotland & Islands	1.06	0.04	[0.99, 1.14]
East England	1.07	0.02	[1.02, 1.12]
Insular Scotland	1.08	0.06	[0.96, 1.21]
Lower North England	1.08	0.03	[1.02, 1.15]
New England	1.08	0.04	[1.00, 1.17]
East Central England	1.09	0.03	[1.03, 1.16]
Scotland	1.10	0.03	[1.04, 1.16]
West Central England	1.11	0.03	[1.04, 1.18]
NYC	1.12	0.04	[1.04, 1.20]
North East England	1.14	0.05	[1.04, 1.26]
Canada (urban)	1.15	0.02	[1.09, 1.21]
Western US	1.15	0.03	[1.09, 1.21]
Canada (rural)	1.17	0.03	[1.12, 1.24]
Ireland	1.17	0.04	[1.07, 1.28]
Philadelphia	1.17	0.02	[1.12, 1.22]
Southern US	1.17	0.03	[1.10, 1.24]
North Midland US	1.18	0.03	[1.11, 1.26]
Northern US	1.18	0.03	[1.11, 1.26]
Wales	1.18	0.03	[1.11, 1.25]
Raleigh US	1.19	0.03	[1.13, 1.26]
South Midland US	1.19	0.03	[1.13, 1.26]
Midwest US	1.20	0.03	[1.14, 1.26]
Northern Cities US	1.24	0.04	[1.15, 1.33]
Rochester NY (AAE)	1.35	0.03	[1.27, 1.44]
Princeville NC (AAE)	1.39	0.03	[1.31, 1.48]
Washington DC (AAE)	1.49	0.02	[1.42, 1.56]

spontaneous speech across other languages may still be smaller than in English and retain cross-linguistic differences akin to those reported by Chen (1970), and thus English would still retain its status as a language with a distinct realization of the VE.

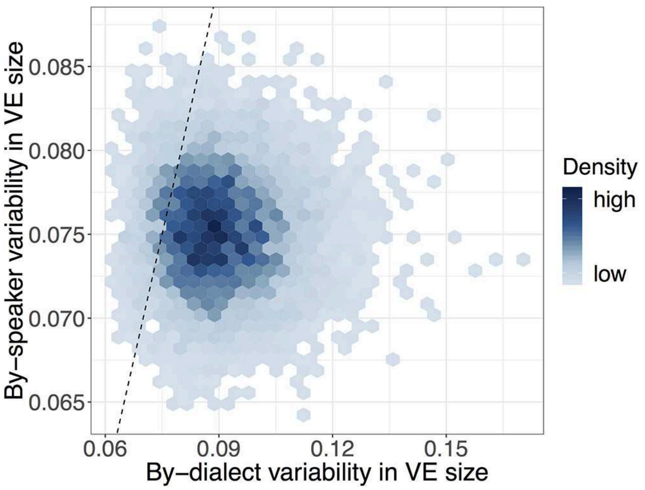


FIGURE 3 | Heatmap of posterior samples of by-dialect ($\hat{\sigma}_{dialect}$) and by-speaker ($\hat{\sigma}_{speaker}$) voicing effect standard deviations. Equal variability is indicated by the dashed line, with darker shades indicating a greater density of samples.

The first research question (section 6.1) considered how the VE was modulated in spontaneous speech, averaging across dialects. To what extent dialects themselves differ in VE was the focus of the second research question. As shown in section 6.2, English was shown to exhibit a range of different VE sizes across individual dialects. The dialects with the smallest and largest VEs—Scottish Englishes and AAE, respectively—were expected to show these values given evidence of additional phonological rules governing vowel duration in these varieties (Aitken, 1981; Holt et al., 2016; Rathcke and Stuart-Smith, 2016; Farrington, 2018). Beyond these varieties, dialects appear to differ gradiently from each other, ranging in VE values from around 1.05 in South West England to 1.24 in the Northern Cities region (**Figure 2**). As opposed there being a single "English" VE value, there appears to be a range of VE sizes within the language. Such a finding further complicates the notion that English has a particular and large VE relative to other languages. Imagining these different dialects as "languages" with minimally different phonological structures, this finding demonstrates that such similar "languages" can have very different phonetic effects (Keating, 1985). This in turn underlies a more nuanced approach to the question of whether English truly differs from other languages in its VE size: not only may English have varieties with greater or lesser VE sizes, but other languages may also exhibit similar dialectal VE ranges.

Individual speakers are also shown to vary in the realization of the VE, though the extent of this variability is rather limited when compared to variability across dialects (**Figure 3**): that is, whilst dialects appear to demonstrate a range of possible VE patterns, individual speakers vary little from their dialect-specific baseline values. Such a finding supports an interpretation where the VE has a dialect-specific value which speakers learn as part of becoming a speaker of that speech community. The limited extent of speaker variability could predict that the VE will be stable within individual English dialects, given the key role of synchronic speaker variability as the basis for sound change (Ohala, 1989; Baker et al., 2011). This would need checking on a dialect-by-dialect basis, however, given recent evidence of Glaswegian undergoing weakening in its vowel duration patterns (Rathcke and Stuart-Smith, 2016). It also highlights the need for studies addressing both synchronic and diachronic variability across dialects, which we hope to address in future work. One important caveat to the finding is that it assumes that all the dialects analyzed in this study contain only speakers who are speakers of that dialect: if a given dialect had a particularly large degree of by-speaker variability, it could be that this could reflect the existence of multiple speakers of different dialects (and thus different VE patterns) within that particular dialect coding. This is unlikely to be a particular problem in this study, however, as a separate model that allows for by-speaker variability to vary on a per-dialect basis showed that no dialect with a sufficiently large number of tokens exhibited overly large by-speaker variability (section 6.2).

By using speech data from multiple sources and multiple dialects, it has been possible to investigate variability of a phonological feature across "English" overall, examine variability at the level of individual dialects and speakers, and reveal the extent of English-wide phonetic variability that was not previously apparent in studies of individual dialects and communities. In this sense, our "large-scale" approach, using consistent measures and controlling factors, enables us to understand the nature of dialectal variability in the English VE directly within the context of both other dialects and English as a whole.

Whilst this kind of study extends the scope of analysis for (socio)phonetic research, there are of course a number of limitations that should be kept in mind in studies of this kind. This study of the English VE predominantly uses data from automatic acoustic measurements, in turn calculated from forced aligned-segmented datasets. All forced-alignment tools have a minimum time resolution (often 10 ms), a minimum segment duration (often 30 ms), and there always exists the possibility of poor or inaccurate alignment. This is a necessary consequence of the volume of data used in this study: there is simply *too much* data to manually check and correct all durations, and so the best means of limiting these effects is through sensible filtering and modeling of the data. For example, segments with aligned durations of less than 50 ms were excluded, since accurately capturing the duration of a vowel this small could be difficult given the time resolution of the aligner. This decision could exaggerate the size of the VE estimation, as only the most reduced vowels have been removed from the data. Another property of forced alignment which impacts our study of VE is that aligners will only apply the phonological segment label to the segment, meaning that it is possible to only examine VE in terms of *phonological* voicing specification (i.e., whether a segment is underlyingly voiced or not), as opposed to whether the segment itself was realized with phonetic voicing. For example, the realization of the stop as devoiced (Farrington, 2018) or as a glottal stop (Smith and Holmes-Elliott, 2018), or the relative duration of the closure preceding the vowel (Lehiste, 1970; Port and Dalby, 1982; Coretta, 2019), could affect VE size which is not controllable by exclusively using phonological segment labels. How this kind of phonetic variation, and the more general relationship between a "phonological" and a "phonetic" VE, should be understood would certainly be an interesting project for future work. Finally, given the diversity of formats and structures of the corpora available for this study, it has only been possible to categorize and study dialects in a rather broad "regional" fashion. Similarly, we were unable to investigate the effect of speaker age due to the heterogenous coding of age across the corpora: we agree this is an important dimension that we have attempted to account for in the approach to statistical modeling, and is certainly necessary to examine in future work. Whilst these limitations may be less suitable for approaching other questions in phonetics and sociolinguistics which are concerned with variability at a more detailed level, the approach taken in this study points to a promising first step toward exposing the structures underlying fine-grained phonetic variability at a larger level across multiple speakers and dialects of a language.

8. CONCLUSION

The recent increase in availability of spoken-language corpora, and development of speech and data processing tools have now made it easier to perform phonetic research at a "large-scale"—incorporating data from multiple different corpora, dialects, and speakers. This study applies this large-scale approach to investigate how the English Voicing Effect (VE) is realized in spontaneous speech, and the extent of its variability across individual dialects and speakers. Little has been known about how the VE varies across dialects bar a handful of studies of specific dialects (Aitken, 1981; Tauberer and Evanini, 2009; Holt et al., 2016). English provides an interesting opportunity to directly examine how phonetic implementation may differ across language varieties with minimally different phonological structures (Keating, 1985). By applying tools for automatic acoustic analysis (McAuliffe et al., 2019) and statistical modeling (Carpenter et al., 2017), it was found that the English VE is substantially smaller in spontaneous speech, as compared with controlled laboratory speech, and is modulated by a range of phonetic factors. English dialects demonstrate a wide degree of variability in VE size beyond that expected from specific dialect patterns such as the SVLR, whilst individual speakers are relatively uniform with respect to their dialect-specific baseline values. In this way, this study provides an example of how large-scale studies can provide new insights into the structure of phonetic variability of English and language more generally.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JT extracted the data, performed the statistical analysis, and wrote the first draft of the manuscript. All authors contributed to the conception and design of the study. All authors contributed to manuscript revision, and read and approved the submitted version.

FUNDING

The research reported here is part of SPeech Across Dialects of English (SPADE): Large-scale digital analysis of a spoken language across space and time (2017–2020); ESRC Grant ES/R003963/1, NSERC/CRSNG Grant RGPDD 501771-16, SSHRC/CRSH Grant 869-2016-0006, NSF Grant SMA-1730479 (Digging into Data/Trans-Atlantic Platform), and was also supported by SSHRC #435-2017-0925 awarded to MS and a Fonds de Recherche du Québec Société et Culture International Internship award granted to JT.

ACKNOWLEDGMENTS

We acknowledge the crucial contribution of The SPADE Consortium (https://spade.glasgow.ac.uk/the-spade-consortium/), which comprises the Data Guardians who generously shared their datasets which are reported in this paper, and/or whose datasets were used in the development of the ISCAN tool. No SPADE research would have been possible without The SPADE Consortium. This research reported in this paper is an extended version of a preliminary report in Tanner et al. (2019). We would like to thank the audiences of the 2019 Montreal-Ottawa-Toronto Phonology Workshop and UK Language Variation & Change 12 for their feedback on previous versions of this research. We also thank SPADE team members, especially Michael Goodale, Rachel Macdonald, and Michael McAuliffe for assistance with data management.

REFERENCES

Aitken, A. J. (1981). *The Scottish Vowel Length Rule. The Middle English Dialect Project*. Edinburgh.

Anderson, J., Beavan, D., and Kay, C. (2007). "The Scottish corpus of texts and speech," in *Creating and Digitizing Language Corpora*, eds J. C. Beal, K. P. Corrigan, and H. L. Moisl (New York, NY: Palgrave), 17–34. doi: 10.1057/9780230223936_2

Baker, A., Archangeli, D., and Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Lang. Variat. Change* 23, 347–374. doi: 10.1017/S0954394511000135

Barr, D. J., Levy, R., Sheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang*. 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Boberg, C. (2018). "Dialects of North American English," in *Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Oxford: John Wiley and Sons), 450–461. doi: 10.1002/9781118827628.ch26

Bois, J. W. D., Chafe, W. L., Meyer, S. A., Thompson, S. A., and Martey, N. (2000). *Santa Barbara Corpus of Spoken American English*. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Brysbaert, M., and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* 41, 977–990. doi: 10.3758/BRM.41.4.977

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R J*. 10, 395–411. doi: 10.32614/RJ-2018-017

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw*. 76, 1–32. doi: 10.18637/jss.v076.i01

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22, 129–159. doi: 10.1159/000259312

Clopper, C. G., Pisoni, D. B., and de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *J. Acoust. Soc. Am*. 118, 1661–1676. doi: 10.1121/1.2000774

Cohen Priva, U., and Gleason, E. (2018). "The role of fast speech in sound change," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 1512–1517.

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). *Audio BNC: The Audio Edition of the Spoken British National Corpus*. Technical report, Oxford. Available online at: http://www.phon.ox.ac.uk/AudioBNC

Coleman, J., Renwick, M. E. L., and Temple, R. A. M. (2016). Probabilistic underspecification in nasal place assimilation. *Phonology* 33, 425–458. doi: 10.1017/S0952675716000208

Coretta, S. (2019). An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. *Glossa* 4, 1–25. doi: 10.5334/gjgl.869

Crystal, T. H., and House, A. S. (1982). Segmental durations in connected speech signals: preliminary results. *J. Acoust. Soc. Am*. 72, 705–716. doi: 10.1121/1.388251

Cuartero, N. (2002). *Voicing assimilation in Catalan and English* (Ph.D. thesis). Universitat Autónoma de Barcelona, Barcelona, Spain.

Denes, P. (1955). Effect of duration on the perception of voicing. *J. Acoust. Soc. Am*. 27, 761–764. doi: 10.1121/1.1908020

Docherty, G. (1992). *The Timing of Voicing in British English Obstruents*. Berlin; New York, NY: Foris. doi: 10.1515/9783110872637

Dodsworth, R. (2013). Retreat from the Southern Vowel Shift in Raleigh, NC: social factors. *Univ. Pennsylvania Work. Pap. Linguist*. 19, 31–40. Available online at: https://repository.upenn.edu/pwpl/vol19/iss2/5/

Dodsworth, R., and Kohn, M. (2012). Urban rejection of the vernacular: the SVS undone. *Lang. Variat. Change* 24, 221–245. doi: 10.1017/S0954394512000105

Ernestus, M., Hanique, I., and Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *J. Phonet*. 38, 60–75. doi: 10.1016/j.wocn.2014.08.001

Fabricius, A. H. (2000). *T-glottalling between stigma and prestige: a sociolinguistic study of Modern RP* (Ph.D. thesis). Copenhagen Business School, Copenhagen, Denmark.

Farrington, C. (2018). Incomplete neutralization in African American English: the cast of final consonant devoicing. *Lang. Variat. Change* 30, 361–383. doi: 10.1017/S0954394518000145

Fromkin, V. A. (1977). "Some questions regarding universal phonetics and phonetic representations," in *Linguistic Studies Offered to Joseph Greenberg on the Occasion of His Sixtieth Birthday*, ed A. Juilland (Saratoga, NY: Anma Libri), 365–380.

Fromont, R., and Hay, J. (2012). "LaBB-CAT: an annotation store," in *Australasian Language Technology Workshop 2012, Vol. 113*, 113–117.

Fruehwald, J. (2013). *The phonological influence on phonetic change* (Ph.D. thesis). University of Pennsylvania, Pennsylvania, PA, United States.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511790942

Geng, C., Turk, A., Scobbie, J. M., Macmartin, C., Hoole, P., Richmond, K., et al. (2013). Recording speech articulation in dialogue: evaluating a synchronized double electromagnetic articulography setup. *J. Phonet*. 41, 421–431. doi: 10.1016/j.wocn.2013.07.002

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Vol. 1* (San Francisco, CA), 517–520. doi: 10.1109/ICASSP.1992.225858

Greenbaum, S., and Nelson, G. (1996). The International Corpus of English (ICE project). *World English*. 15, 3–15. doi: 10.1111/j.1467-971X.1996.tb00088.x

Guy, G. (1980). "Variation in the group and the individual: the case of final stop deletion," in *Locating Language in Time and Space*, ed W. Labov (New York, NY: Academic Press), 1–36.

Harris, M., and Umeda, N. (1974). Effect of speaking mode on temporal factors in speech: vowel duration. *J. Acoust. Soc. Am*. 56, 1016–1018. doi: 10.1121/1.1903366

Heffner, R.-M. S. (1937). Notes on the lengths of vowels. *Am. Speech* 12, 128–134. doi: 10.2307/452621

Hewlett, N., Matthews, B., and Scobbie, J. M. (1999). "Vowel duration in Scottish English speaking children," in *Proceedings of 14th The International Congress of Phonetic Sciences* (San Francisco, CA).

Hibbitt, G. W. (1948). *Diphthongs in American speech: a study of the duration of diphthongs in the contextual speech of two hundred and ten male undergraduates* (Ph.D. thesis). Columbia University, New York, NY, United States.

Holmes-Elliott, S. (2015). *London calling: assessing the spread of metropolitan features in the southeast* (Ph.D. thesis). University of Glasgow, Glasgow, Scotland.

Holt, Y. F., Jacewicz, E., and Fox, R. A. (2016). Temporal variation in African American English: the distinctive use of vowel duration. *J. Phonet. Audiol.* 2. doi: 10.4172/2471-9455.1000121

House, A. S. (1961). On vowel duration in English. *J. Acoust. Soc. Am.* 33, 1174–1178. doi: 10.1121/1.1908941

House, A. S., and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* 25, 105–113. doi: 10.1121/1.1906982

Jacewicz, E., and Fox, R. A. (2013). "Cross-dialectal differences in dynamic formant patterns in American English vowels," in *Vowel Inherent Spectral Change*, eds G. S. Morrison and P. F. Assmann (Berlin: Springer), 177–198. doi: 10.1007/978-3-642-14209-3_8

Jones, D. (1948). *An Outline of English Phonetics*. New York, NY: E. P. Dutton & Company.

Keating, P. (1984). Phonetic and phonological representation of stop consonant voicing. *Language* 60, 189–218. doi: 10.2307/413642

Keating, P. (2006). "Phonetic encoding of prosodic structure," in *Speech Production: Models, Phonetic Processes, and Techniques*, eds J. Harrington and M. Tabain (New York, NY: Psychology Press), 197–186.

Keating, P. A. (1985). "Universal phonetics and the organization of grammars," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed V. A. Fromkin (New York, NY: Academic Press), 115–132.

Kendall, T., and Farrington, C. (2018). *The Corpus of Regional African American Language. Version 2018.10.06*, Eugene, OR.

Kenyon, J. S. (1940). *American Pronunciation*. Ann Arbor, MI: George Wahr.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *J. Acoust. Soc. Am.* 54, 1102–1104. doi: 10.1121/1.1914322

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J. Acoust. Soc. Am.* 59, 1208–1221. doi: 10.1121/1.380986

Kleinschmidt, D. F. (2018). Structure in talker variability: how much is there and how much can it help? *Lang. Cogn. Neurosci.* 34, 1–26. doi: 10.1080/23273798.2018.1500698

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110167467

Labov, W., Cohen, P., Robins, C., and Lewis, J. (1968). *A Study of the Non-Standard English of Negro and Puerto Rican Speakers in New York City*. Technical Report 1 & 2, Linguistics Laboratory, University of Pennsylvania.

Labov, W., and Rosenfelder, I. (2011). "New tools and methods for very large scale measurements of very large corpora," in *New Tools and Methods for Very-Large-Scale Phonetics Research Workshop*, Pennsylvania, PA.

Labov, W., Rosenfelder, I., and Fruehwalf, J. (2013). One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89, 30–65. doi: 10.1353/lan.2013.0015

Lehiste, I. (1970). Temporal organization of higher-level linguistic units. *J. Acoust. Soc. Am.* 48:111. doi: 10.1121/1.1974906

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* 100, 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Liberman, M. (2018). Corpus phonetics. *Annu. Rev. Linguist.* 5, 91–107. doi: 10.1146/annurev-linguistics-011516-033830

Liberman, M. A., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the h&h theory," in *Speech Production and Speech Modelling*, Vol. 4 of NATO ASI Series, eds W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer Academic Publishers), 403–439. doi: 10.1007/978-94-009-2037-8_16

Lisker, L. (1957). Linguistic segments, acoustic segments and synthetic speech. *Language* 33, 370–374. doi: 10.2307/411159

Lisker, L. (1985). The pursuit of invariance in speech signals. *J. Acoust. Soc. Am.* 77, 1199-1202. doi: 10.1121/1.392185

Luce, P. A., and Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *J. Acoust. Soc. Am.* 78, 1949–1957. doi: 10.1121/1.392651

Mack, M. (1982). Voicing-dependent vowel duration in English and French: monolingual and bilingual production. *J. Acoust. Soc. Am.* 71, 173–178. doi: 10.1121/1.387344

McAuliffe, M., Coles, A., Goodale, M., Mihuc, S., Wagner, M., Stuart-Smith, J., et al. (2019). "ISCAN: A system for integrated phonetic analyses across speech corpora," in *Proceedings of the 19th International Congress of Phonetic Sciences* (Melbourne, VIC).

McAuliffe, M., Scolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017a). *Montreal Forced Aligner [computer program]*. Available online at: https://montrealcorpustools.github.io/Montreal-Forced-Aligner/

McAuliffe, M., Stengel-Eskin, E., Socolof, M., and Sonderegger, M. (2017b). "Polyglot and Speech Corpus Tools: a system for representing, integrating, and querying speech corpora," in *Proceedings of Interspeech 2017* (Stockholm). doi: 10.21437/Interspeech.2017-1390

Nicenboim, B., and Vasishth, S. (2016). Statistical methods for linguistic research: foundational ideas - part II. *Lang. Linguist. Compass* 10, 591–613. doi: 10.1111/lnc3.12207

Ohala, J. (1989). "Sound change is drawn from a pool of synchronic variation," in *Language Change: Contributions to the Study of Its Causes*, eds L. E. Breivik and E. H. Jahr (Berlin: Mouton de Gruyter), 173–198.

Peterson, G. E., and Lehiste, I. (1960). Duration of syllable nuclei in English. *J. Acoust. Soc. Am.* 32, 693–703. doi: 10.1121/1.1908183

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). *Buckeye Corpus of Spontaneous Speech, 2nd Edn*. Columbus, OH: Ohio State University.

Port, R. F., and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Percept. Psychophys.* 32, 141–152. doi: 10.3758/BF03204273

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *J. Acoust. Soc. Am.* 51, 1296–1303. doi: 10.1121/1.1912974

Rathcke, T., and Stuart-Smith, J. (2016). On the tail of the Scottish Vowel Length Rule in Glasgow. *Lang. Speech* 59, 404–430. doi: 10.1177/0023830915611428

Roettger, T. B., Winter, B., and Baayen, R. H. (2019). Emergent data analysis in phonetic sciences: towards pluralism and reproducibility. *J. Phonet.* 73, 1–7. doi: 10.1016/j.wocn.2018.12.001

Rosen, N., and Skriver, C. (2015). Vowel patterning of Mormons in Southern Alberta, Canada. *Lang. Commun.* 42, 104–115. doi: 10.1016/j.langcom.2014.12.007

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., et al. (2014). *FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281*.

Rositzke, H. A. (1939). Vowel-length in General American speech. *Am. Speech* 15, 99–109. doi: 10.2307/408728

Smith, J., and Holmes-Elliott, S. (2018). The unstoppable glottal: tracking rapid change in an iconic British variable. *English Lang. Linguist.* 22, 323–355. doi: 10.1017/S1360674316000459

Solanki, V. J. (2017). *Brains in dialogue: investigating accommodation in live conversational speech for both speech and EEG data* (Ph.D. thesis). University of Glasgow, Glasgow, Scotland.

Solé, M.-J. (2007). "Controlled and mechanical properties in speech," in *Experimental Approaches to Phonology*, eds P. Beddor and M. Ohala (Oxford: Oxford University Press), 302–321.

Sonderegger, M., Bane, M., and Graff, P. (2017). The medium-term dynamics of accents on reality television. *Language* 93, 598–640. doi: 10.1353/lan.2017.0038

Sonderegger, M., Stuart-Smith, J., McAuliffe, M., Macdonald, R., and Kendall, T. (2019). "Managing data for integrated speech corpus analysis in SPeech Across Dialects of English (SPADE)," in *Open Handbook of Linguistic Data Management*, eds A. Berez-Kroeker, B. McDonnell, E. Koller, and L. Collister (Cambridge: MIT Press).

Stuart-Smith, J., Jose, B., Rathcke, T., MacDonald, R., and Lawson, E. (2017). "Changing sounds in a changing city: an acoustic phonetic investigation of real-time change over a century of Glaswegian," in *Language and a Sense of Place: Studies in Language and Region*, eds C. Montgomery and E. Moore (Cambridge: Cambridge University Press), 38–65. doi: 10.1017/9781316162477.004

Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Lab. Phonol.* 6, 505–549. doi: 10.1515/lp-2015-0015

Summers, W. V. (1987). Effects of stress and final consonant voicing on vowel production: articulatory and acoustic analyses. *J. Acoust. Soc. Am*. 82, 847–863. doi: 10.1121/1.395284

Sweet, H. (1880). *A Handbook of Phonetics*. London: MacMillan & Co.

Tagliamonte, S., and Temple, R. (2005). New perspectives on an ol variable: (t, d) in British English. *Lang. Variat. Change* 17, 281–302. doi: 10.1017/S0954394505050118

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of York English: was/were variation as a case study for statistical practice. *Lang. Variat. Change* 24, 135–178. doi: 10.1017/S0954394512000129

Tanner, J., Sonderegger, M., Stuart-Smith, J., and SPADE-Consortium (2019). Vowel duration and the voicing effect across English dialects. *Univers. Toronto Work. Pap. Linguist*. 41, 1–13. doi: 10.33137/twpl.v41i1.32769

Tauberer, J., and Evanini, K. (2009). "Intrinsic vowel duration and the post-vocalic voicing effect: some evidence from dialects of North American English," in *Proceedings of Interspeech*.

Thomas, C. K. (1947). *An Introduction to the Phonetics of American English*. New York, NY: Ronald Press Company.

Thomas, E. R. (2001). *An Acoustic Analysis of Vowel Variation in New World English*. American Dialect Society.

Trudgill, P. (1999). *The Dialects of England*. Oxford: Blackwell.

Umeda, N. (1975). Vowel duration in American English. *J. Acoust. Soc. Am*. 58, 434–445. doi: 10.1121/1.380688

Vasishth, S., Nicenboim, B., Beckman, M., Li, F., and Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: a tutorial introduction. *J. Phonet*. 71, 147–161. doi: 10.1016/j.wocn.2018.07.008

Yuan, J., and Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Commun*. 65, 67–74. doi: 10.1016/j.specom.2014.06.001

Yuan, J., Liberman, M., and Cieri, C. (2006). "Towards an integrated understanding of speaking rate in conversation," in *Proceedings of Interspeech 2006*, Pittsburgh, PA.

Yuan, J., Liberman, M., and Cieri, C. (2007). "Towards an integrated understanding of speech overlaps in conversation," in *Proceedings of the International Congress of Phonetic Sciences XVI* (Saarbrücken), 1337–1340.

Zimmerman, S. A., and Sapon, S. M. (1958). Note on vowel duration seen cross-linguistically. *J. Acoust. Soc. Am*. 30, 152–153. doi: 10.1121/1.1909521

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 May 2020
doi: 10.3389/frai.2020.00035

Implicit Standardization in a Minority Language Community: Real-Time Syntactic Change among Hasidic Yiddish Writers

*Isaac L. Bleaman**

Department of Linguistics, University of California, Berkeley, Berkeley, CA, United States

The recent turn to "big data" from social media corpora has enabled sociolinguists to investigate patterns of language variation and change at unprecedented scales. However, research in this paradigm has been slow to address variable phenomena in minority languages, where data scarcity and the absence of computational tools (e.g., taggers, parsers) often present significant barriers to entry. This article analyzes socio-syntactic variation in one minority language variety, Hasidic Yiddish, focusing on a variable for which tokens can be identified in raw text using purely morphological criteria. In non-finite particle verbs, the overt tense marker *tsu* (cf. English *to*, German *zu*) is variably realized either between the preverbal particle and verb (e.g., *oyf-tsu-es-n* up-to-eat-INF 'to eat up'; the conservative variant) or before both elements (*tsu oyf-es-n* to up-eat-INF; the innovative variant). Nearly 38,000 tokens of non-finite particle verbs were extracted from the popular Hasidic Yiddish discussion forum *Kave Shtiebel* (the 'coffee room'; kaveshtiebel.com). A mixed-effects regression analysis reveals that despite a forum-wide favoring effect for the innovative variant, users favor the conservative variant the longer their accounts remain open and active. This process of rapid implicit standardization is supported by ethnographic evidence highlighting the spread of language norms among Hasidic writers on the internet, most of whom did not have the opportunity to express themselves in written Yiddish prior to the advent of social media.

Keywords: corpus sociolinguistics, minority languages, syntactic variation, particle verbs, standardization, Yiddish, Hasidic Jews

OPEN ACCESS

Edited by:
James N. Stanford,
Dartmouth College, United States

Reviewed by:
Julie A. Van Dyke,
Haskins Laboratories, United States
Kathryn H. Franich,
University of Delaware, United States

****Correspondence:***
Isaac L. Bleaman
bleaman@berkeley.edu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *25 November 2019*
Accepted: *21 April 2020*
Published: *29 May 2020*

Citation:
Bleaman IL (2020) Implicit Standardization in a Minority Language Community: Real-Time Syntactic Change among Hasidic Yiddish Writers. Front. Artif. Intell. 3:35. doi: 10.3389/frai.2020.00035

1. INTRODUCTION

In recent years, sociolinguists have increasingly turned to social media platforms like Twitter to investigate large-scale patterns of language variation and change. Some of the areas that have been addressed include gender and style (Bamman et al., 2014), the geographic diffusion of lexical variants (Eisenstein et al., 2014; Huang et al., 2016; Grieve et al., 2018), and the grammatical and social constraints on orthographic variation (Eisenstein, 2015). Social media corpora have increased not only the number of speakers (or writers) whose data can be analyzed in a single research project, but also the range of variables that can be effectively studied: in a corpus containing tens of millions or even billions of words, one can uncover robust sociolinguistic patterns even for variables that occur with low frequency in conversational interviews.

While the field of sociolinguistics continues to gain valuable insights from "big data" in social media, most of this work contributes to our understanding of only a handful of language varieties—American English chief among them. The research bias favoring monolingual majority communities has been a longstanding problem in sociolinguistics (Meyerhoff and Nagy, 2008; Stanford, 2016; Guy and Adli, 2019), and it certainly extends to social media studies. Unfortunately, many of the existing tools in computational linguistics (including stemmers, part-of-speech taggers, and syntactic parsers) were not designed to support minority language data. Even if raw text data can be obtained—which is not always the case, especially for endangered varieties—the lack of computational tools to process the data presents fundamental challenges to large-scale research on these languages and their users. This may explain why social media studies of minority languages, including Welsh (Jones et al., 2013), Māori (Keegan et al., 2015), Limburgish, and Frisian (Nguyen et al., 2015), tend to focus on macro-level social phenomena such as language choice rather than micro-level linguistic phenomena such as grammatical variation.

One minority language that has been considered exemplary of "resource-poor" languages is Yiddish (Genzel et al., 2009), which is spoken at home by some 170,000 Americans, 86% of whom reside in New York State (U.S. Census Bureau, 2015). According to the engineers who developed Google Translate in Yiddish, the reason for this designation is the problem of data scarcity: the lack of large parallel corpora makes it difficult to obtain the training data necessary for automatic machine translation. They argue that if engineers can overcome these challenges for Yiddish, they would be well-positioned to address similar challenges in other "low-resource" languages—"a very important public service that will help preserve these languages and make literature in these languages available to the rest of the world" (Genzel et al., 2009, p. 6).

Ironically, the availability of Google Translate in Yiddish has led to the proliferation of fake Yiddish websites, thus exacerbating the problem of data scarcity for other applications. For example, students interested in the usage of particular words and phrases must now sift through pages of search results containing both reliable Yiddish-language sources, including newspaper articles, and unreliable ones, including blogs whose authors used Google Translate to render their posts in many different languages, presumably to increase reader traffic[1]. For linguists interested in the grammar of minority languages, including Yiddish, the ubiquity of machine-translated text raises serious questions about the reliability of data taken from the internet. For example, software like BootCaT (Baroni and Bernardini, 2004), which builds corpora by scraping the web for pages containing target-language keywords, inadvertently includes some of these machine-translated websites. Fortunately, recent years have also seen an increase in the number of *real* Yiddish websites, including discussion forums designed for Hasidic Jews who make up the vast majority of today's native speakers.

The goal of this article is to show not only that a corpus study using online Hasidic Yiddish is feasible, but also that it can yield novel findings about linguistic variation comparable to those obtained from social media studies of majority languages like English. The current study analyzes socio-syntactic variation on a popular Hasidic Yiddish discussion forum, focusing on particle verbs and the relative position of the non-finite tense marker *tsu* 'to.' Tokens of this variable can be identified in raw text using purely morphological criteria, without the need for a part-of-speech tagger, a parser, or even a dictionary, none of which have yet been developed for Hasidic Yiddish. In addition to linguistic constraints on the variable, the study uncovers a significant social fact: although the discussion forum shows a modest increase in the probability of the innovative variant, users favor the conservative variant the longer their accounts remain open and active. This finding, framed as an example of *rapid implicit standardization* on the internet, is supported by ethnographic evidence highlighting the role of the discussion forum in spreading language norms among its Hasidic Jewish users.

This study has important consequences for the analysis of variation in minority languages, as it demonstrates the utility of computational methods even for a language variety, Hasidic Yiddish, without an extensive online presence or linguistically processed corpora of any size[2]. Given that majority languages including English are actually *over-represented* on large social media platforms like Twitter (Mocanu et al., 2013), it is especially encouraging that smaller discussion forums can provide adequate minority language data for variationist sociolinguistics. This study also contributes to our understanding of contemporary Hasidic Yiddish, which has been overshadowed in linguistic research by projects focused on the European dialects spoken before the Holocaust (Nove, 2018). The results of this study corroborate the view—one taken for granted by sociolinguists but still uncommon among specialists in Yiddish studies—that seemingly inconsistent and disorderly linguistic behavior among Hasidic Jews is in fact principled and orderly, conditioned by linguistic and extra-linguistic factors in predictable ways.

The article is organized as follows. Section 2 introduces the online community (the discussion forum *Kave Shtiebel*) from which a sociolinguistic corpus was built for this study. Evidence will be presented to show that these anonymous writers are Hasidic Jews who reside primarily in New York. Section 3 introduces the syntactic variable, which has not previously been mentioned in linguistic descriptions of Yiddish; for this reason, most of the hypotheses about quantitative constraints (presented in 3.2) are drawn from studies of particle verb phrases in English, which involve a different set of variants. Section 4 describes the method for automatically extracting tokens of the variable from the forum's posts. Section 5 presents the results of the statistical

[1] As of November 2019, the first page of Google search results for the high-frequency trigram *vos iz der* 'what is the.MASC.SG' includes a webpage entitled *vos iz der taytsh fun lebn?* 'what is the meaning of life?' from GotQuestions.org, an evangelical Christian missionary blog. Each page has been machine-translated into dozens of languages, and the Yiddish version is nearly incomprehensible.

[2] The largest annotated corpus available for any variety of Yiddish is the *Corpus of Modern Yiddish* (http://web-corpora.net/YNC/search/), a 4-million-word collection limited to texts published in the standardized YIVO orthography (YIVO, 1999), which is not used in any Hasidic community. By contrast, the *Kave Shtiebel* corpus assembled for this study contains approximately 29 million words from Hasidic Yiddish writers.

analysis of the variation, laying out the relevant constraints and their interpretations. This section also offers a detailed discussion of two seemingly contradictory effects relating to real-time syntactic change among forum users (presented in 5.2). Finally, section 6 summarizes the conclusions and the questions they raise for future sociolinguistic studies of minority language corpora.

2. THE CORPUS AND THE COMMUNITY

For the religiously conservative Hasidic community, the maintenance of a Jewish vernacular language reflects a broader ideology that opposes acculturation to non-Jewish norms (Isaacs, 1999). Hasidic Jews in the United States constitute an urban speech community, as they are geographically concentrated in a few Yiddish-speaking neighborhoods in Brooklyn and Upstate New York. Yiddish is used as a medium of instruction in private Hasidic schools, which are segregated by gender and feature very different curricula in terms of both content and language. Boys receive an essentially monolingual education in Yiddish; English is only taught from third to eighth grade (approximately age 7–13), and during those years, it is only taught for ninety minutes a day in the very late afternoon, a period reserved for all non-religious subjects. Girls, by contrast, have a fully bilingual curriculum from first grade through the end of high school, with Yiddish used for religious subjects and English for secular subjects (Fader, 2009, pp. 22–23). The imbalance in bilingual proficiency between men and women has been cited by community members as one reason why Yiddish-language discussion forums tend to be men's spaces. By contrast, the most popular forum among Hasidic women, imamother.com, is written in English.

While the Hasidic community is committed to the maintenance of Yiddish, its leaders do not support efforts to standardize the language. The use of Yiddish is strictly enforced in Hasidic schools, but subjects like "grammar" (norms of language use) and "composition" (writing skills) are viewed as distractions from serious religious study and are not emphasized in Hasidic curricula. Hasidic Jews have played virtually no role in the standardization efforts of secular organizations like the YIVO and the League for Yiddish, and Hasidic publishers have never endorsed their standards. This is not to say that Hasidic Jews lack standard language ideologies; as mentioned below in section 3, Hasidic consultants agree that in non-finite particle verbs, one variant often sounds "more correct" than the other. The language ideologies of Hasidic men and women are discussed in more depth in Bleaman, 2018.

Universal literacy in Yiddish means that Hasidic newspapers and magazines enjoy sizable readerships, but very few Hasidic adults have a regular need to write in Yiddish after finishing school. This was articulated to me offline in a sociolinguistic interview I conducted with Berl (33 years old; Monsey, NY), who works as a freelance writer. (All names of interviewees are pseudonyms.)

> It used to be, until… literally ten or fifteen years ago, if a person wasn't a Yiddish writer and he wasn't studying in *koylel* [religious school for married men] where he'd have to write down his ideas about the Torah or take notes… there literally wasn't, that kind of person didn't have to write a single sentence in Yiddish in twenty years. There was nowhere to write, no reason to write, nobody to write for. At work he'd write in English, obviously, nobody writes in Yiddish at work. His grocery list is English. He just didn't write. Zero.
> (Translated from Yiddish.)

Berl's reference to "ten or fifteen years ago" alludes to the advent of Hasidic blogs, and later of online discussion forums and WhatsApp groups specifically for Hasidic users—all of which have afforded community members new opportunities to express themselves in written Yiddish. The role of the internet in rejuvenating Hasidic writing was articulated in many of the sociolinguistic interviews I conducted with Hasidic Jews offline (Bleaman, 2018). Another Hasidic man, Duvid (36; Monsey), told me that before participating in *Kave Shtiebel*'s poetry competition he had never done any creative writing whatsoever, in Yiddish or any other language.

Hasidic discussion forums have existed since at least 2005. In that year, a now-defunct Hebrew-language forum called *Hyde Park* had a Yiddish-language subforum called *heymishe shtusim* 'Hasidic nonsense.' The subforum was designed as a place where Yiddish-speaking Hasidic men could post their questions and concerns related to sexual matters (masturbation, premature ejaculation, marital relations) which are considered taboo to discuss publicly. Over time, writers began to discuss other more mundane topics, including sports, which are also seen as inappropriate for Hasidic Jews. In 2006, a standalone forum called *iVelt* (short for *idishe velt* 'Jewish world'; ivelt.com/forum) was launched, which has since become increasingly mainstream in its ultra-Orthodox religious and social outlook.

A second independent forum, *Kave Shtiebel* (kaveshtiebel.com), was launched in February 2012. Its name refers to the 'coffee room' of a study or prayer house, where men can take a break and chat casually over a cup of coffee. *Kave Shtiebel* (KS) was founded in response to mounting frustration with the moderation of *iVelt*, where posts that were critical of Hasidic power structures (especially the authority of the rabbis) were routinely deleted. KS prides itself on giving writers the freedom to post socially critical content, alongside other topics including history, science, religion, politics, and poetry. This commitment is codified in its guidelines for new members. In recent years, KS users have also come together to publish an *offline* magazine, with original content touching on religious and secular topics. This magazine, *Veker* 'lit., one who awakens,' is sold on Amazon and at newsstands in Brooklyn and other neighborhoods.

Because the users of Hasidic discussion forums are largely inexperienced amateur writers—having attended schools where writing skills are not developed systematically—there is understandably a significant amount of variation in the written Yiddish found on the internet today, including orthographic inconsistencies. At the same time, one might expect the overall amount of variation to decrease over time, as writers develop their skills and acquire norms from one another. Indeed, there is anecdotal evidence suggesting this trend. A lively conversation ensued in response to a message I recently posted to KS (November 10, 2019) soliciting specific examples of writing conventions that users have acquired since joining the forum. The responses mentioned norms in spelling and punctuation, such as the difference between a comma and a period. One user, writing under the username *Gefilte fish*, identified the singular role that KS has played in his development as a writer:

> *Kave Shtiebel* taught me not only how to write in Yiddish, spelling, grammar, but I couldn't even use the Hebrew keyboard before I got here. Here I've learned how to spell in Yiddish, including the difference between *in* and *and*, and many other things that I can't recall at the moment. Go back to my first posts from 2012 and you'll see that I spelled like a grandma. (Grandmas, don't take it personally. You write very well. I mean no disrespect, it's just a turn of phrase.) [. . .] Of my graduating class in *yeshiva* [religious school] I couldn't name even three people who can write a "sentence" (*zats*?) in any language, not Yiddish, not English, not Hebrew.
> (Translated from Yiddish.)

Gefilte fish's inexperience as a writer prior to joining KS is indicated by his having acquired the ability to type in Hebrew (Yiddish is written using Hebrew characters) and the orthographic distinction between two basic function words (*in* and *and*, which are spelled differently in Hasidic publications but are homophonous in the Central Yiddish dialect used by Hasidic Jews: [ɪn]). The quote also suggests that his development as a writer is ongoing: he questions whether *zats* is the correct Yiddish word for 'sentence,' which he initially presents as an English borrowing in Hebrew characters.

Another user, *Katle kanye*[3], wrote that whenever he isn't sure which spelling or vocabulary variant to use, he types the options into KS's search box to compare their relative frequencies. If neither variant is more common than the other, he opts for the one used by the KS writers whom he most respects.

The current study provides quantitative support—from one area of Yiddish syntax, non-finite particle verbs—showing that KS writers are shifting toward greater use of normative grammatical features over time as they interact on the forum. This is a process that I term *rapid implicit standardization*, and it will be explicated in the discussion that follows.

2.1. The "Coffee Room" and Its Hasidic Writers

The linguistic data for this study come from the Hasidic discussion forum *Kave Shtiebel*. In order to use an online forum to analyze variation in a minority language variety, it is important to establish who its users are and to what speech community they belong offline. The fact that nearly all KS writers are Hasidic men from the greater New York area is clear from the language of the forum itself: KS is written in Yiddish following Hasidic orthographic conventions, and its posts regularly include phrases from rabbinic texts written in Hebrew and Aramaic (which are the core of Hasidic boys' but not girls' education) as well as borrowings from New York English. Not surprisingly, some of the most active threads are concerned with politics and current events in the New York Hasidic community (and satellite towns such as Lakewood, NJ).

KS is extremely protective of users' confidentiality, and users virtually never disclose any personal information in their profiles. Still, it is possible to identify broad demographic trends in the forum's metadata. The founders of KS granted me access to the database containing all public posts, which I downloaded most recently on October 23, 2019. (This same content could have been obtained by scraping the forum's pages.) The corpus, representing approximately seven and a half years of activity, contains 29 million word tokens across 392,660 posts by 2,194 users.

Figure 1 plots all the posts in the database, grouped by the day of the week on which they were written and binned into hourly intervals (Eastern Time Zone). The figure reveals two important social facts: First, KS writers are concentrated on the East Coast, since there is a daily lull in activity when East Coast residents typically sleep. Second, virtually all KS writers observe the Jewish Sabbath from Friday evening through Saturday evening, when the use of computers and smartphones is prohibited. The expectation that users observe the Sabbath is also mentioned in KS's guidelines for new users. Tellingly, its Yiddish localization of the forum software phpBB translates "Saturday" as *motse-shabes* 'the evening following the Sabbath,' which assumes that all posts with a "Saturday" timestamp are written after sunset.

The same trend of Orthodox religious observance is evident from a plot of all posts to KS during the Jewish month of Tishrei, coinciding with parts of September and October (**Figure 2**). Virtually no messages are posted during the major holidays (Rosh Hashanah, Yom Kippur, etc.) when the use of electronic devices is prohibited.

While the two graphs suggest that KS users are Orthodox Jews on the East Coast, they do not show that users are necessarily Hasidic New Yorkers. The only direct evidence of this comes from offline interactions with KS users. I first joined KS as a way to recruit Hasidic Jews for sociolinguistic interviews as part of a larger research project (Bleaman, 2018). Although my Yiddish recruitment letter did not specify demographic criteria for participation, the 12 KS users I met in person had remarkably uniform social characteristics. All of them were native Yiddish-speaking men, aged 25–36, and affiliated with Hasidic communities—most from the Satmar community, but with some representation from the Vizhnitz and Tosh communities. All of them were living in Hasidic neighborhoods in the New York area (Williamsburg, Boro Park, and Monsey), had attended Hasidic schools for their entire education, had gone through arranged marriages, and were working for Hasidic businesses.

Although this discussion strongly suggests that KS writers belong to the Hasidic Yiddish speech community offline, it would be a mistake to draw any definitive conclusions about

[3] *Katle kanye* is the most well-known Hasidic blogger, and his reputation extends far beyond KS. His self-published book on the problems of Hasidic education was recently recognized by the *Forward 50*: https://forward.com/series/forward-50/2018/katle-kanye/.

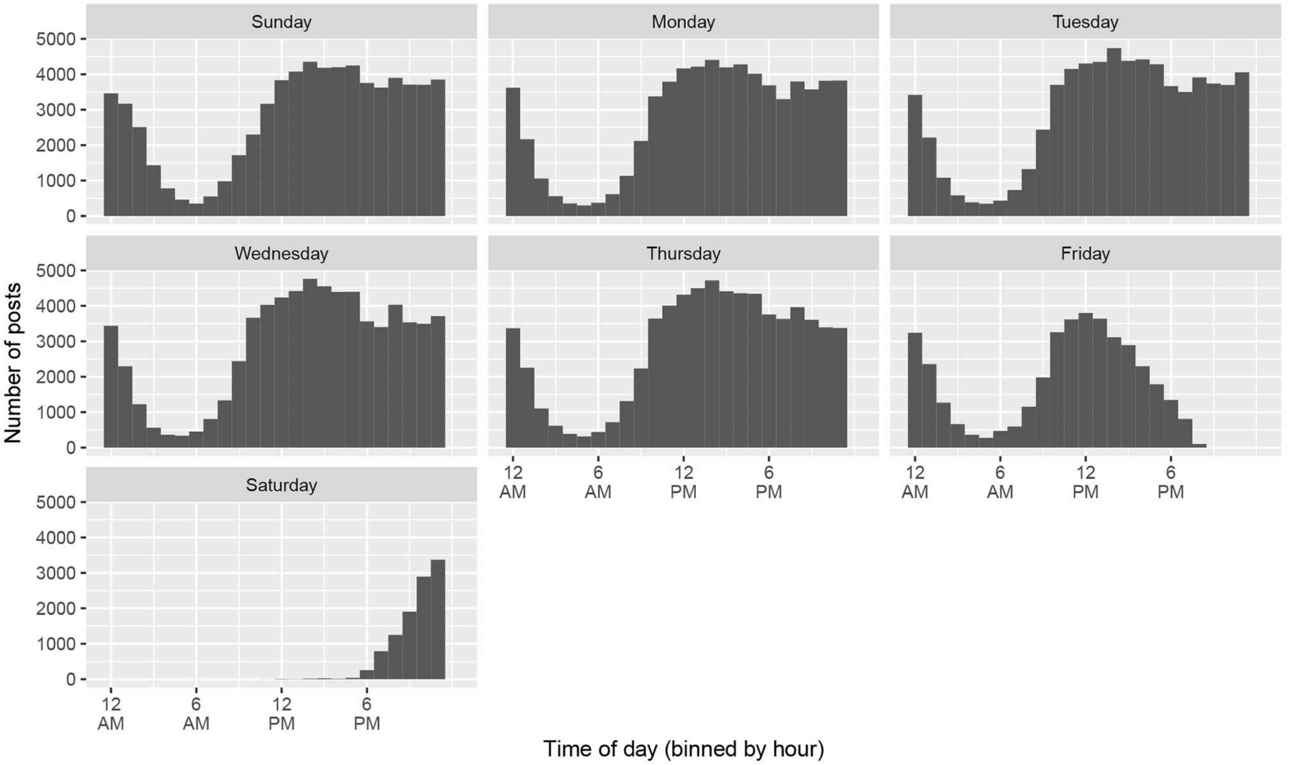


FIGURE 1 | All posts from *Kave Shtiebel* by timestamp and day of the week (Eastern Time Zone).

"(Hasidic) Yiddish" as a whole based on a study of the forum alone. Doing so would overlook the inherent stylistic differences that exist between spoken and written language, as well as the possibility of internet- or even platform-specific registers of written language. Some research in computational sociolinguistics has found that social media writing approximates certain aspects of speech, such as the high frequency of first- and second-person pronouns compared to third-person pronouns in discussion groups (Yates, 1996, pp. 40–42) and the linguistic constraints on orthographic *t,d*-deletion (e.g., *lef* for *left*) and *g*-deletion (*talkin*) on Twitter (Eisenstein, 2015). However, other studies have shown that online registers make use of features (or rates of features) that diverge from users' spoken repertoires, such the use of African-American English variants by gay white Reddit users from the UK (Ilbury, 2019) or the use of restrictive relative clauses headed by a pronoun (e.g., *we who #FeelTheBern*), which are readily found on Twitter despite being stylistically marked (Conrod et al., 2016). The mixed results of these studies should caution us against extrapolating linguistic patterns in speech from linguistic behavior in writing on the internet.

The comparability of speech and online writing is further complicated for contemporary Yiddish, due to the opposition of Hasidic leadership to online communication. Hasidic rabbis have issued decrees against the use of internet-enabled smartphones (Deutsch, 2009), and Hasidic Jews who require internet access for work are expected to install community-mandated web filters (Fader, 2017). One of the ways this is enforced is that parents must certify in writing that they have installed filters on their phones (making them "kosher") before they can enroll their children in school. These filters block access to websites that are considered improper for Hasidic visitors; some evidently even block *Kave Shtiebel*, although not *iVelt*. Despite these prohibitions—and as the impetus for these prohibitions—Hasidic Jews are increasingly using the internet for everyday communication and entertainment. Just as Hasidic entrepreneurs have realized the potential of the internet for business (Deutsch, 2009, p. 4), so too have everyday Hasidic consumers become avid users of internet media, circulated on Hasidic websites and in Hasidic WhatsApp groups.

These considerations highlight some of the limitations of KS data. Not only does the forum reflect the online writing of men of a narrow age range, but its users engage in practices that are considered subversive by the standards of the Hasidic community. Still, KS is one of the most well-known Yiddish websites, Hasidic or otherwise, and its members come from the largest community of Yiddish speakers in the United States. There is also no clear evidence suggesting that the language of KS differs radically from written Hasidic Yiddish offline, especially in its grammatical properties. Even if the results of a study of KS cannot directly address language patterns in the wider speech community, they may offer insights which can become the hypotheses for further research.

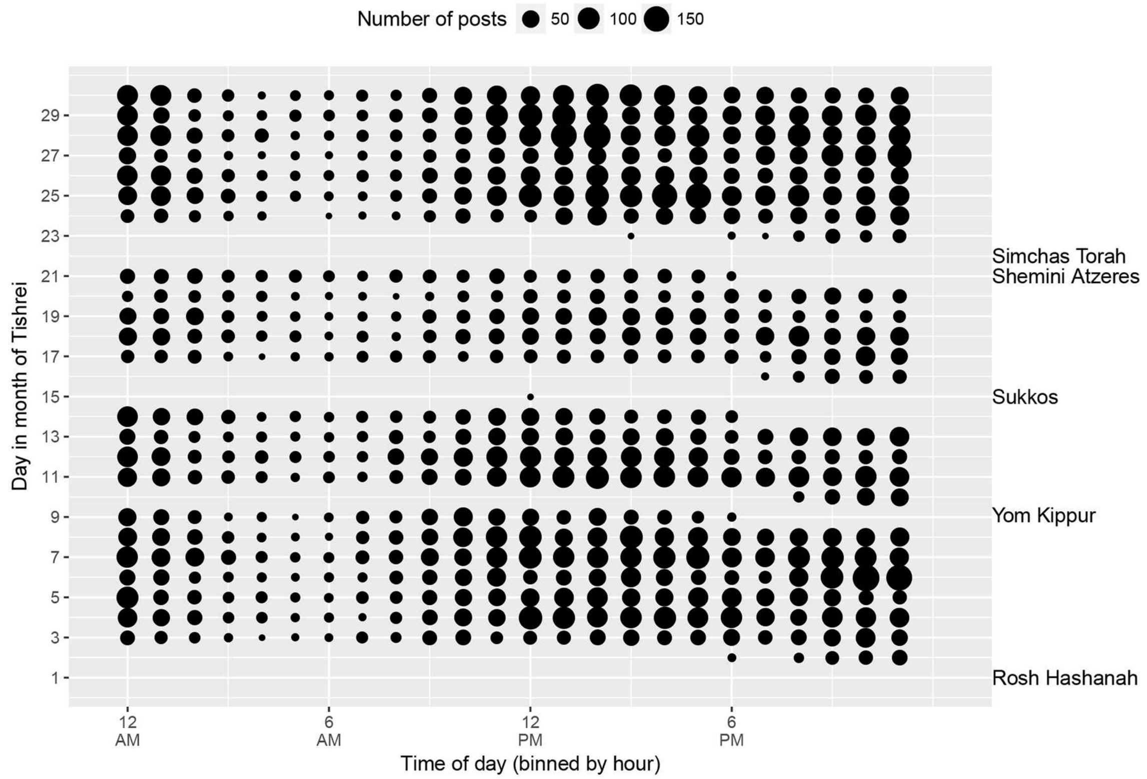


FIGURE 2 | All KS posts written during the Jewish month of Tishrei, 5773-5780 (2012-2019), by time of day (Eastern Time Zone) and day of the month. Jewish holidays when computer use is prohibited are indicated to the right of the plot.

3. PARTICLE VERB VARIATION IN YIDDISH

The linguistic focus of this study is a syntactic alternation involving particle verbs in non-finite tense phrases in Yiddish.

Particle verbs (also known as *phrasal verbs*) are combinations of verbs and preposition- or adverb-like particles, which together form a close semantic unit (Dehé, 2015, p. 611). In English, particles invariably appear after the verb (e.g., *throw up*, *hang out*). In Yiddish, particles appear before the verb in most syntactic contexts. For example, particles always precede the verb in the infinitive, such as when a particle verb phrase appears as the complement of a modal like *must*:

(1) damols vel ikh muzn **uf-es-n** nokh a por tatsn kugl.
then will I must **up-eat-INF** more a few trays kugel
'Then I'll have to **eat up** a few more trays of kugel [Sabbath casserole].' (September 8, 2016)

(Note: Yiddish is written in the Hebrew alphabet. All examples from the KS corpus are provided in standard YIVO transliteration. Hyphens have been added to show morpheme boundaries.)

While modals select for bare infinitival verb phrase (VP) complements, other verbal, nominal, and adjectival predicates select for tense phrase (TP) complements. This context licenses an overt non-finite tense marker, *tsu* (a cognate of English *to* and German *zu*), in addition to the infinitival suffix on the verb (*-n*). The contrast between non-finite VP and TP complements is illustrated below in (2) and (3); note that the contrast is also found in English.

(2) er muz (***tsu**) es-n.
he must (***to**) eat-INF
'He must (***to**) eat.'

(3) a. er hot probirt **tsu** es-n.
he has tried **to** eat-INF
'He tried **to** eat.'
b. s'iz tsayt **tsu** es-n.
it's time **to** eat-INF
'It's time **to** eat.'
c. ... kedey **tsu** es-n.
in.order **to** eat-INF
'... in order **to** eat.'

3.1. Variable Word Order in Non-finite Particle Verbs

The variation analyzed in this article concerns the relative position of *tsu* 'to' in non-finite particle verbs. Generally, *tsu* appears between the preverbal particle and the verb, and the combination is usually written as a single word (e.g., *oyf-**tsu**-es-n* up-**to**-eat-INF 'to eat up'). However, *tsu* sometimes appears before both the preverbal particle and the verb, usually separated by a space (e.g., ***tsu** oyf-es-n* **to** up-eat-INF). Examples of the two variants are shown below in (4) and in (5). These sets of near-minimal pairs are both from the KS corpus.

(4) a. ikh hob nisht probirt **oys-tsu-rekhen-en** di
I have not tried **out-to-calculate-INF** the
mayles fun yedn mentsh.
virtues of every person
'I wasn't trying **to enumerate** the virtues of every person.' (October 8, 2013)

b. ikh gey afile nisht probir-n **tsu oys-rekhen-en**
I go even not try-INF **to out-calculate-INF**
di mayles derfun.
the virtues of.that
'I'm not even going to try **to enumerate** the virtues of that.' (August 4, 2015)

(5) a. shoyn tsayt **oyf-tsu-her-n** mit di kinderishe
already time **up-to-hear-INF** with the childish
zakhn.
things
'It's time **to stop** with all these childish things.' (June 22, 2016)

b. shoyn tsayt **tsu oyf-her-n** mit di narishe un
already time **to up-hear-INF** with the stupid and
zinloze mehalekh.
senseless approach
'It's time **to stop** with this stupid, senseless approach.' (August 19, 2014)

Throughout this article, the label *PtoV* (**P**article-**to**-**V**erb) will be used to refer to the variant in which *tsu* 'to' intervenes between the particle and verb, as in (4-a) and (5-a). The label *toPV* (**to P**article-**V**erb) will be used when *tsu* precedes both elements, as in (4-b) and (5-b).

The *PtoV* order is the only possibility mentioned in the Yiddish grammatical literature (Mark, 1978, p. 330; Schaechter, 1995, p. 64) and the only one taught in university-level Yiddish classes. It is also by far the more common variant in contemporary Hasidic Yiddish, as this article will show. The use of *toPV* is very likely to be a change in progress: It is relatively rare in publications printed in pre-Holocaust Eastern Europe[4] and it is not attested in the dialectological data on the Hasidic community's European source dialects[5]. Many non-Hasidic native speakers of Yiddish judge *toPV* to be totally ungrammatical. Nevertheless, the *toPV* order is readily found in informal Hasidic Yiddish text on the web and is also attested in newer Hasidic publications indexed in Google Books.

As with other proposed syntactic variables, one must ask whether *PtoV* and *toPV* are truly variants of one another—that is, whether they are equivalent either in meaning or in discourse function. The existence of near-minimal pairs like (4) and (5) may be the best evidence of functional equivalence. As a secondary check, three native speakers of Hasidic Yiddish (all *Kave Shtiebel* users) were asked to comment on a number of example sentences. When shown sentences with one variant, native speakers informed me that the other variant would "mean the same thing" (but that *PtoV* often sounded more "correct"). Of course, while these intuitions suggest equivalence, native speakers are likely to be unaware of, or unable to characterize, the various factors that correlate with the use of either variant (see Silverstein, 1981). It is one task of variationist analysis to determine what these factors might be.

Since Yiddish grammars do not mention the *toPV* variant, the factors that affect the use of *PtoV* or *toPV* are not at all understood. Fortunately, the variable lends itself to analysis using a social media corpus like KS, for a few different reasons. First, non-finite particle verbs do not occur very frequently in spoken Yiddish, so a very large corpus is required to obtain the requisite number of tokens for thorough analysis[6]. Second, tokens of the variable can be identified on purely morphological grounds, simply by extracting all strings beginning with a valid Yiddish particle and ending with the infinitival suffix *-n*, with *tsu* appearing either before or after the particle. Using morphological criteria to identify tokens is particularly helpful in the case of Hasidic Yiddish, a minority language variety in which there are no dictionaries or part-of-speech taggers to rely on when searching through raw text.

3.2. Particle Verb Variation in English and Predictions for Yiddish

The variable word order of particle verb phrases is among the most well-studied alternations in the syntactic literature. In English, the variation involves the relative ordering of postverbal particles and non-pronominal objects in transitive verb phrases, as shown in (6).

(6) a. He **looked up** the information.
b. He **looked** the information **up**. (Dehé, 2002, pp. 3–4)

When discussing the variation in English, I follow the convention of Dehé (2002) who uses the term "continuous" to refer to instances when the verb and particle are adjacent (6-a) and "discontinuous" when they are not (6-b).

[4] There are a few lexicalized exceptions for which *toPV* is common (*iber-zets-n* 'translate,' *iber-tsayg-n* 'convince,' *iber-rash-n* 'surprise,' *unter-drik-n* 'oppress') in which the "particle" variably behaves like a prefix, so that it is not stressed and does not separate from the verbal root in the past participle or present tense conjugations. These exceptions are likely borrowings from Modern German, since *iber* 'over' and *unter* 'under' are not productive prefixes in Yiddish.
Simon Neuberg (pers. comm.) has sent me examples of the *toPV* order that he has encountered in modern literary sources. Most of them come from Soviet writers whose native dialect was Northern ("Litvish") Yiddish, which is geographically and linguistically distant from the Transcarpathian varieties considered to be the source of contemporary Hasidic Yiddish.

[5] I am grateful to Lea Schäfer and her student assistant Marc Brode, both of the *Syntax of Eastern Yiddish Dialects* (SEYD) project, for processing the relevant data from the *Language and Culture Atlas of Ashkenazic Jewry*. The *Atlas*'s survey includes question 020.020/021 'it's not nice *to stick out* your tongue,' which explicitly targets the acceptability of *toPV*. The only informants who supplied or accepted *toPV* came from cities and towns in present-day Belarus, Lithuania, and Eastern Poland. Transcarpathian speakers rejected *toPV*.

[6] Even in the longest sociolinguistic interviews I conducted in the New York area, it is difficult to find more than five tokens of non-finite particle verbs per speaker.

Although the syntactic alternation in Hasidic Yiddish (pronouncing *tsu* 'to' before or after the preverbal particle) differs from the alternation in English (pronouncing the object before or after the postverbal particle), they are superficially similar in that one variant involves strict adjacency between verb and particle while the other does not. In other words, *toPV* could be described as "continuous" because the verb and particle are adjacent, and *PtoV* could be described as "discontinuous" because the verb and particle are separated by *tsu*. For this reason, it is worth considering the literature on particle verb variation in English in order to formulate hypotheses about the variation in Yiddish, which has not been documented before[7].

In one of the earliest sociolinguistic studies of the alternation, Kroch and Small (1978) identify the "degree of semantic dependence of particle on the verb" as one linguistic predictor of the word order variation. The intuition is that combinations of verb and particle whose meaning cannot be predicted from the sum of their parts (e.g., *throw up* 'vomit,' *put up* 'temporarily house') function as standalone predicates and are most easily parsed when the verb and particle are adjacent. The idiomaticity of the particle verb combination has been shown in many studies to be among the strongest predictors of the variation, and considerable work has been done to define it formally (see Lohse et al., 2004; Bannard, 2005). The tendency for idiomatic combinations to remain structurally or linearly adjacent is also involved in categorical grammaticality judgments. Zeller (2001, pp. 89–90) observes that German allows for the topicalization of particles when the combination is semantically transparent [e.g., *auf-geh-en* up-go-INF 'rise' in (7-a)] but not when it is idiomatic [e.g., *auf-hör-en* up-hear-INF 'stop' in (7-b)]. The same judgments hold for English (8) and Yiddish (9).

(7) a. **Auf** geht die Sonne im Osten (aber **unter** geht sie im Westen).
up goes the sun in.the east (but **down** goes it in.the west)
'The sun rises in the east (but sets in the west).'

b. ***Auf** hat Peter mit dem Trinken gehört.
up has Peter with the drinking heard
'Peter stopped drinking.'

(8) a. And **out** they went.

b. *And **out** they made.
(intended: 'kissed passionately')

(9) a. **arop** iz er gefaln.
downward is he fallen
(*arop-fal-n* downward-fall-INF 'fall down')

'DOWN he fell (not OVER).'

b. ***op** iz dos ayz nit gegangen.
down is the ice not gone
(*op-gey-n* down-go-INF 'thaw, defrost')

'The ice didn't THAW.' (Diesing, 1997, p. 384)

Gries (2001) presents an overview of various factors that linguists have proposed as predictors of the variation in English and offers a unified account based on processing effort/cost: for example, the more morphosyntactically complex an object is (correlated with the number of words it contains), the more difficult or cognitively "expensive" it is to process the discontinuous order. The same holds true of idiomatic particle verb combinations compared to ones that are semantically transparent. If speakers aim to facilitate effective communication by minimizing the processing cost for the listener, then it follows that more complex particle verb phrases (e.g., long idiomatic ones) will favor the continuous order, which is cognitively easier to process (Rohdenburg, 1996). A related proposal has been offered by Lohse et al. (2004), who focus on the size of the processing domain and its relationship to the syntactic and semantic properties of the particle verb construction.

In Yiddish, as in English and other Germanic languages, combinations of particle and verb vary in terms of their semantic transparency or compositionality (Mark, 1978, p. 308; Diesing, 1997, pp. 383–384; Talmy, 2000, p. 297). Directional particles combine with motion verbs to yield semantically transparent combinations (e.g., *aroys-gey-n* outward-walk-INF 'walk out, exit'). By contrast, non-directional particles combined with the same verbs often have idiomatic meanings (*oys-gey-n* out-walk-INF 'expire; die')[8]. If idiomatic combinations prefer to remain adjacent (*toPV*), it could be because they are (variably) derived via the morphological incorporation of the particle into the verb; this would (variably) prevent the intervention of *tsu* between the two elements, just as it prevents the topicalization of the particle (Diesing, 1997, p. 384). Under this theory, these particles would behave (at least some of the time) like genuine prefixes, which are always adjacent to their verbs (*toP*[*refix*]*V*; see Biskup et al., 2011 on prefix and particle verbs in German). Regardless of how semantic transparency is reflected in syntactic derivations, its role will be examined in the current study by means of grouping Yiddish particles into different types, discussed in section 4.2.

Another predictor of the variation in English is the information entropy of the particle, which is used to gauge its productivity or ability to associate with different verbs (Schnoebelen, 2008). Information entropy works in this way: For each particle, we generate a list of all of the unique verbs with which it appears in the corpus, and the number of times it appears with each of those verbs. Entropy is low if a particle only appears with a small number of different verbs, and high if it appears with a variety of verbs at roughly equal rates[9]. It is assumed that particles with low entropy are less productive than high entropy

[7] The use of a different set of labels for the variants in Yiddish (*toPV* and *PtoV*, rather than "continuous" and "discontinuous") is justified for several reasons. First, the English variants both have postverbal particles, while the Yiddish variants both have preverbal particles. Second, in English the continuous variant is evaluated as more normatively "correct" than the discontinuous variant, while in Yiddish the "discontinuous" or nonadjacent variant (*PtoV*) is preferred. Finally, using new labels minimizes the confusion likely to arise when referring to a "continuous variant" in syntax and a "continuous variable" in statistical analysis.

[8] Mark (1978) and others have observed that the non-directional particles often contribute some aspectual meaning to the verb (as in English *eat* vs. *eat up*). See also Gold, 1999, chapter 1.

[9] The entropy of a particle is defined as the negative sum of the probability of each unique verb that occurs with that particle multiplied by the log probability of that

particles. Combinations with low productivity particles may be considered more "wordlike," and are expected to favor the variant in which the particle and verb are adjacent: the continuous order in English, and *toPV* in Yiddish.

Social factors have also been shown to condition the variation in English. Kroch and Small (1978) demonstrate that talk radio hosts use the continuous order at a significantly higher rate than listeners do when calling into the show. They take this as evidence that the standard language ideology favoring the continuous order is active in everyday linguistic behavior and can serve as a marker of status[10]. Haddican and Johnson (2012) find significant differences between UK/Irish English and North American English, with the latter favoring the continuous order at higher rates than the former in both production (gleaned from Twitter data) and perception (a sentence rating task). They also find that the relative frequency of the discontinuous order has increased over time, based on evidence drawn from a historical corpus.

If standard language ideology promotes the *PtoV* variant in Yiddish, then one might hypothesize a positive correlation between *toPV* and the use of other non-standard features, including non-standard spellings. To test this hypothesis, the analysis below will consider whether there is a non-standard orthographic form anywhere in the non-finite particle verb token (in the particle, in the verb, or in the use of *tsi* for *tsu*, a common spelling variant reflecting the spoken dialect of Hasidic Jews).

Finally, if *toPV* is a change in progress within Hasidic Yiddish, then we also expect younger speakers (and writers) to use the innovative *toPV* variant at higher rates than older speakers (and writers). Unfortunately, KS cannot currently be used to analyze age-based sociolinguistic stratification, because the corpus represents less than eight years of activity (February 2012 through October 2019) and because its writers seem to come mostly from the same generational cohort (married men under 40). However, KS can still be used to study the effect of time, on the forum as a whole and in the posts of individual users. The hypotheses with regard to syntactic change in progress are presented in section 4.2.

4. DATA PROCESSING AND ANALYSIS

4.1. Building the Dataset: Extracting Tokens of Non-finite Particle Verbs

On October 23, 2019, the database containing all public posts from KS was downloaded and imported into a data frame, with one column representing the content of the post and other columns containing the post's metadata. Using Python scripts, each message was stripped of HTML tags and text quoted from other users, and then tokenized—i.e., converted from a long text string to a list of individual words, excluding punctuation. Each token was also stripped of all characters not contained in the standard Hebrew alphabet, e.g., apostrophes and diacritics, including those found within pre-combined Unicode characters sometimes used in non-Hasidic Yiddish. Word-final letter forms (*langer nun*, *shlos-mem*, etc.) were also converted to non-final forms to avoid certain inconsistencies within Hasidic orthography[11].

At this point, Yiddish grammars (in particular, Mark, 1978, pp. 301–311 and Jacobs, 2005, p. 210) were consulted to generate a list of all Yiddish particles[12], supplemented by common variants used in Hasidic Yiddish[13]. Posts were then searched for all word strings beginning with any of these particles, followed by *tsu* (or *tsi*, a dialect spelling), and ending with the infinitival suffix, *-n*. In this way, it was possible to rely on morphological criteria to identify particle verbs, rather than a pre-defined dictionary. This yielded a list of 36,370 *potential* examples of *PtoV* non-finite particle verbs, representing 3,704 unique strings.

These potential *PtoV* tokens were used to generate a list of all potential verbs, i.e., just the substring after the particle and *tsu*. This list of potential verbs—containing exactly 1,300 unique strings—was exported to a text file and hand-checked for accuracy[14]. A number of these items were removed because they were not actually verbs[15], and additional non-standard spellings were added to the list. A script was then used to assemble the full list of all theoretically possible particle verbs, by combining every particle with every (hand-verified) verb. At this point, all KS posts were searched for matches of all non-finite particle verbs appearing in either order: *PtoV* or *toPV*[16].

This method of using morphological criteria (plus manual verification) to identify non-finite particle verbs yielded 37,858 tokens of either *PtoV* or *toPV*. Of these, 5,553 tokens (14.7%) were of the innovative/non-standard *toPV* variant. This final

same verb, i.e., $H(\text{particle}) = -\sum_{\text{v}\in\text{verbs}} p(\text{v}) \times \log_2 p(\text{v})$. Its implementation in Python is given in Bird et al., 2009, example 6–8.

[10] The standard ideology motivating this difference is unclear. The continuous order may be favored if speakers are reanalyzing particles as prepositions, which according to the norms of standard usage should not appear in sentence-final position.

[11] The removal of final forms (and apostrophes) makes it easier to recognize variant spellings as instances of a single lexical item. For example, verbs with roots derived from Hebrew, e.g., *kholem-en* dream(Hebrew)-INF(Germanic) 'dream,' are inconsistently spelled with final letters (and apostrophes) before the Germanic infinitival suffix. The stripping of final forms in Yiddish is somewhat analogous to converting capital letters to lowercase in English.

[12] Three particles were omitted in order to avoid false positives and categorization errors: *tsu*, which coincides with the non-finite tense marker 'to' and therefore makes the *PtoV* and *toPV* orders indistinguishable; *for*, which is spelled just like the (inseparable) verbal prefix *far*; and *um*, which coincides with the adjectival prefix meaning 'un-.'

[13] Examples include variants that reflect regional pronunciations (e.g., *inter* for *unter* 'under') and reduced forms (e.g., *tsam*, equivalent to standard *tsuzamen* 'together').

[14] A silent letter *alef* is used in Yiddish to break up three adjacent repeated letters (the so-called *mekhitse-alef* 'barrier *alef*'; Katz, 1993, p. 139), as when a single *vov* representing [u] appears in front of a double *vov* representing [v]: *aroys-ts**u***-ALEF-***v**ayz-n* outward-to-show-INF 'to display.' All *mekhitse-alef*s at the beginning of verb strings were manually removed.

[15] For example, the single letter '*g*' appeared in the list of verbs. This is because, purely by coincidence, the plural noun *oystsugn* 'excerpts' looks like it begins with the particle *oys* 'out,' followed by *tsu*, and ending with *-n* (which is also the plural suffix). "*G*" is not a verb (or a word) and was removed from the list.

[16] Because non-finite particle verbs can theoretically appear as one, two, or three separate words with spaces, only the following forms were counted as tokens: one word (*PtoV* or *toPV*), two words (*to PV*), or three words (*to P V*). The other possible spacing patterns were excluded to avoid false positives.

dataset represents 1,768 unique (spelling-normalized) particle verb combinations from 1,165 users.

4.2. Coding Independent Linguistic and Non-linguistic Factors

Each token of the dependent variable (*PtoV* vs. *toPV*) was coded for a variety of potential conditioning factors including social, grammatical, and cognitive predictors (Tamminga et al., 2016). These factors, which were tested in the full regression model, were:

Categorical fixed effects

i. particle type (*directional*, *cognate*, *other*);
ii. whether the verb is an English borrowing (e.g., *arayn-tsu-**sken**-en* inward-to-scan-INF 'to scan in'; *aroys-tsu-**sayn**-en* outward-to-sign-INF 'to sign out');
iii. whether the post has been "liked" by another user;
iv. whether the token contains a non-standard spelling (of particle, verb, or *tsi* for *tsu*);
v. persistence (the variant used most recently within the same post: *PtoV*, *toPV*, or *none*);

Continuous fixed effects

vi. the information entropy of the particle;
vii. the number of phonological segments in the (spelling-normalized) particle verb combination;
viii. the log frequency of the (spelling-normalized) particle verb combination;
ix. the number of days elapsed from user registration to the current post's timestamp (i.e., the user's seniority);
x. the number of days elapsed from the launch of KS to the current post's timestamp (i.e., the age of the forum);

Random effects

xi. writer (username); and
xii. word (spelling-normalized particle verb combination).

The motivation for including some of these factors was presented in section 3.2, along with predictions based on studies of particle verb variation in English. For clarity, the remainder of this subsection will summarize the predictions for all of these factors in order.

The first factor, particle type, is a way to approximate the semantic transparency of the particle verb combination. As noted by Talmy (2000, pp. 297–298), Yiddish particles can be categorized into three distinct types. The first type includes directional particles (e.g., *arayn* 'inward,' *aroys* 'outward,' *aroyf* 'upward,' etc.) that attach freely to all motion verbs, verbs of transfer, etc., and usually contribute a concrete or metaphorical directional reading to the resulting particle verb. Yiddish also has a series of what I call 'cognate' particles, which look like the directional particles but without the initial *ar-* (i.e., *ayn*, *oys*, *oyf*, etc.). These are often translated into English as prepositions ('in,' 'out,' 'up,' etc.) and their semantic contribution is generally more idiosyncratic (e.g., *oys-gey-n* out-go-INF 'expire; die'). The remaining Yiddish particles were classified as "other." Examples of each of the three particle types are shown in **Table 1**. (Note that my labels "directional," "cognate," and "other" correspond to Talmy's (2000, pp. 297–298) terms "long doublet," "short doublet," and "singlet"). If particle verbs with directional particles are maximally transparent in meaning, then perhaps speakers/writers will more readily tolerate their separation from the verb by the presence of intervening *tsu* (i.e., *PtoV*)—much in the same way that Yiddish allows for their topicalization to the front of the sentence (Diesing, 1997, p. 384). If particle verbs with cognate particles are the least semantically transparent, then these combinations should favor strict adjacency (*toPV*). Particles in the catchall "other" category should favor neither variant.

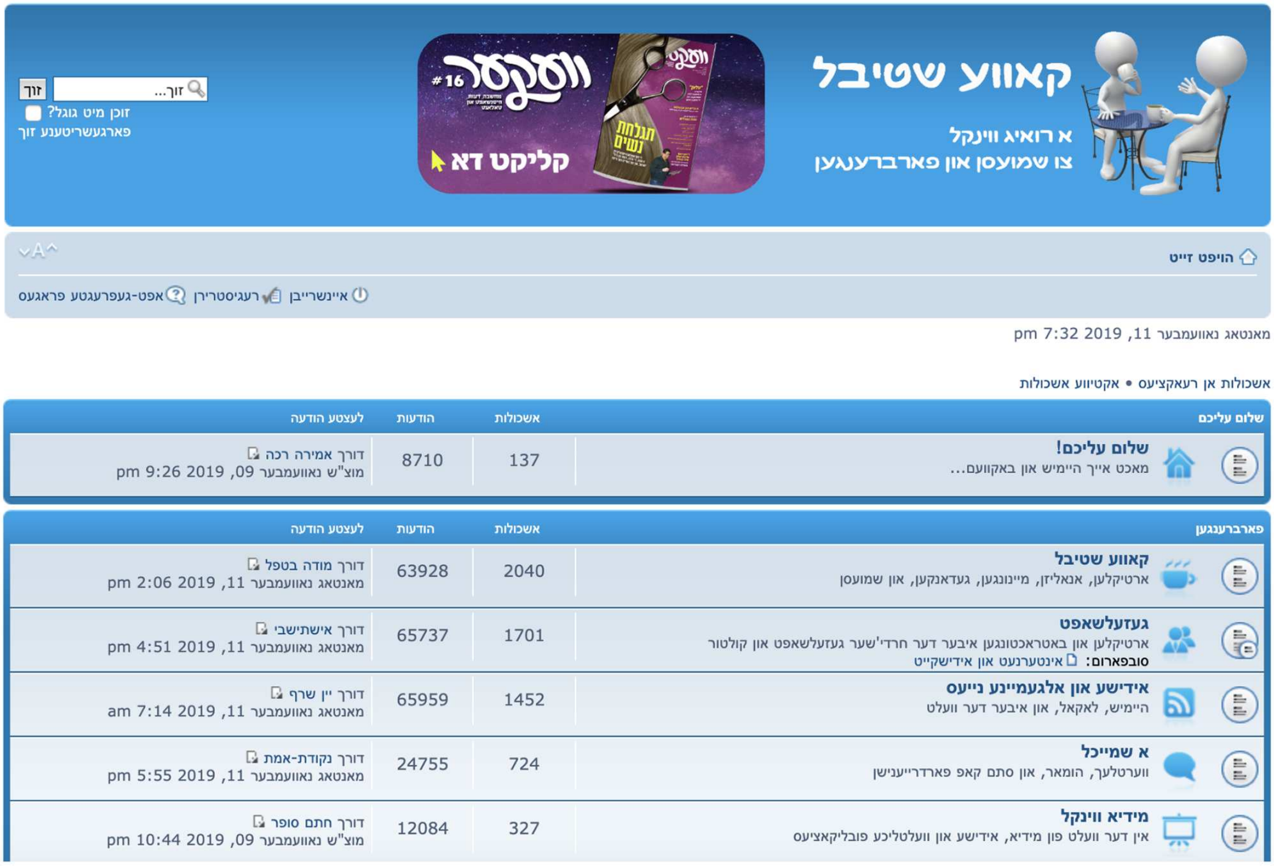
The inclusion of binary factors for whether the verb is an English borrowing, whether the post has been "liked" by another user, and whether the token contains a non-standard spelling is meant to capture intuitions about the social nature of the *toPV* variant. If a writer borrows a particular English particle verb (in which *to* always precedes the verb and particle: ***to** sign in*), we might also expect him to use the innovative/non-standard variant in which *tsu* is the first element (*toPV*; ***tsu** arayn-sayn-en*). Posts that receive a positive social evaluation, in the form of a "like" from another user, might correlate with the use of standard grammatical features, like *PtoV*. Finally, the use of a non-standard spelling in the particle verb token might favor the use of the non-standard variant (*toPV*).

Persistence describes the tendency for tokens of a recently produced variant to influence subsequent tokens of the variable (Scherre, 2001; Tamminga, 2016; see also Weiner and Labov, 1983, p. 47). Some of the effect is due to the fact that the initial token is "drawn from the same distribution" as subsequent tokens (Tamminga, 2016, p. 343), i.e., from the same speaker, who may be biased to produce one variant at a higher or lower rate than the population mean. However, persistence has been found to be significant even in regression models with random effects for speaker, suggesting a more general cognitive basis (Tamminga, 2016). Although persistence is most relevant in spontaneous speech, it has been found to be a significant predictor of particle verb variation even in written corpora (Gries, 2005). Because KS is designed to be a place for casual anonymous conversation (*a ruig vinkl tsu shmuesn* 'a relaxed spot to converse,' as its masthead states; see **Figure 3**), some of the cognitive constraints on speech production may be preserved in this genre of informal writing, as well. Persistence was captured in this study by means of a discrete variable coded for the most recently used variant within the same post (*PtoV*, *toPV*, or *none* if the current token is the first of its post). If writers are biased to repeat tokens within posts, then a previous occurrence of *PtoV* should favor the repetition of *PtoV*, *toPV* should favor repetition of *toPV*, and the first or only token in a particular post (*none*) should not favor either variant.

The information entropy of the particle is meant to capture its productivity. If a particle appears rather predictably only with a small number of different verbs (i.e., low information entropy), the resulting combinations may be more "wordlike" and thus likelier to remain adjacent (*toPV*). Particles with high information entropy attach to a greater variety of different verbs, and the resulting combinations may be less "wordlike" and easier to separate (*PtoV*).

TABLE 1 | Examples of the three particle types.

Particle type	Example particles	Example combination	Translation
directional	*aroys* 'outward,' *aroyf* 'upward'	*aroys-fir-n* outward-lead-INF	'lead out(side)'
cognate	*oys* 'out,' *oyf* 'up'	*oys-fir-n* out-lead-INF	'execute; conclude'
other	*mit* 'with,' *nokh* 'after'	*mit-fil-n* with-feel-INF	'empathize'



מאנטאג נאוועמבער 11, 2019 7:32 pm

אשכולות אן רעאקציעס • אקטיווע אשכולות

שלום עליכם	אשכולות	הודעות	לעצטע הודעה
שלום עליכם! מאכט אייך היימיש און באקוועם...	137	8710	דורך אמירה רכה מוצ"ש נאוועמבער 09, 2019 9:26 pm

פארברענגען	אשכולות	הודעות	לעצטע הודעה
קאווע שטיבל ארטיקלען, אנאליזן, מיינונגען, געדאנקען, און שמועסן	2040	63928	דורך מודה בטפל מאנטאג נאוועמבער 11, 2019 2:06 pm
געזעלשאפט ארטיקלען און באטראכטונגען איבער דער חרדי'שער געזעלשאפט און קולטור סובפארום: אינטערנעט און אידישקייט	1701	65737	דורך אישתישבי מאנטאג נאוועמבער 11, 2019 4:51 pm
אידישע און אלגעמיינע נייעס היימיש, לאקאל, און איבער דער וועלט	1452	65959	דורך יין שרף מאנטאג נאוועמבער 11, 2019 7:14 am
א שמייכל ווערטלעך, הומאר, און סתם קאפ פארדרייעניש	724	24755	דורך נקודת-אמת מאנטאג נאוועמבער 11, 2019 5:55 pm
מידיא ווינקל אין דער וועלט פון מידיא, אידישע און וועלטליכע פובליקאציעס	327	12084	דורך חתם סופר מוצ"ש נאוועמבער 09, 2019 10:44 pm

FIGURE 3 | The front page of *Kave Shtiebel* (screenshot from November 11, 2019). Image published with permission of forum moderators.

The analysis also includes a factor for the number of phonological segments in the (spelling-normalized) particle verb combination. When KS writers were asked to provide judgments on *PtoV*∼*toPV* minimal pairs, some remarked that inserting *tsu* between the particle and verb would make the word "too long" or unwieldy to write and read. Since *PtoV* is usually written as one word but *toPV* as two (i.e., *to PV*), longer particle verb combinations might favor *toPV* merely by virtue of their being longer strings. This hypothesis isn't motivated by existing literature, but rather by users speaking from their personal experience typing on their computers and smartphones. (Note that the number of phonological segments in the string usually coincides with the number of orthographic characters.)

It has been argued in the literature on exemplar models of linguistic knowledge that frequency of occurrence affects the way forms are cognitively stored and produced (e.g., Bybee, 2002). However, the role of frequency in constraining syntactic variation (as opposed to phonological variation) has not been consistent across studies. Some evidence suggests that high lexical frequency can amplify the effects of other constraints but may not have an independent effect of its own (Erker and Guy, 2012). However, attempts at replication have found that constraint effects may actually be stronger for *lower* frequency items (Bayley et al., 2013). The working hypothesis for this study is that since *PtoV* is the overwhelmingly preferred variant (all else being equal), more frequent combinations of particle and verb are likelier to have a larger sheer number of *PtoV* tokens than *toPV* tokens, and therefore a more robust representation of *PtoV* exemplars stored in speakers' episodic memory. Consequently, it is predicted that higher frequency particle verb combinations will favor *PtoV*. Since no standalone corpora of Hasidic Yiddish exist, frequency information for each particle verb combination

TABLE 2 | The five most frequent particle verb combinations from the dataset containing all non-finite particle verb tokens.

Particle verb combo.	Gloss	Translation	Frequency
aroys-breng-en	outward-bring-INF	'bring out; express'	868
on-kum-en	on-come-INF	'arrive'	833
on-nem-en	on-take-INF	'accept'	751
on-heyb-n	on-lift-INF	'start'	618
arayn-gey-n	inward-go-INF	'walk in, enter'	583

was calculated from within the generated dataset of non-finite particle verb tokens. Frequency was based on spelling-normalized combinations of particle and verb, to abstract over any typographical differences in raw tokens. **Table 2** shows the most frequent combinations in the dataset.

The number of days elapsed since user registration (i.e., a given user's seniority on KS at the time of the post) and the number of days elapsed since the launch of KS (i.e., the age of the forum at the time of the post) are meant to capture syntactic change in progress. If users are implicitly acquiring grammatical norms over time as they write and engage with other KS members, there should be a positive correlation between user seniority and the use of *PtoV*. If *toPV* is innovative, then we might expect to find a higher probability of *toPV* over time on the forum as a whole, irrespective of any tendency for individual writers to become more standard. Such an effect, if found, should be very modest, since there is no reason to believe that the user demographics of KS (including age) have shifted much from 2012 to 2019.

Finally, the model includes random intercepts for writer (username) and dictionary word (spelling-normalized particle verb combination), as well as by-writer random slopes for all predictors of interest. The inclusion of random effects is important to account for the inherent variability across individual writers and words. For example, some KS users are also professional writers and editors, and they may inherently favor *PtoV* more than other users, show less sensitivity to word length, etc. There will also inevitably be certain particle verb combinations (such as *op-deyt-n*, which is the English borrowing 'update') that have an atypical baseline rate for the variable (*tsu op-deyt-n* 'to update' is used much more often than *op-tsu-deyt-n*, although both are found in the corpus). Including random effects in the statistical model controls for some of these inherent differences.

5. RESULTS

5.1. Statistical Analysis

The variation in word order (*PtoV* vs. *toPV*) across all 37,858 non-finite particle verb tokens was modeled through logistic mixed-effects regression using the R package lme4 (version 1.1-17; Bates et al., 2015). The fixed effects included in the full model were the factors numbered (i) through (x) in the previous section. All continuous predictors were standardized. The model also included random intercepts for writer (1,165 different usernames) and for word (1,768 different particle verb combinations), and by-writer random slopes (uncorrelated) for all fixed effect terms.

The model's fixed effects are summarized in **Table 3**. *P*-values were calculated based on asymptotic Wald tests. The McFadden's pseudo r^2 for this model was 0.259. Note that a more parsimonious model, excluding all non-significant fixed effects and corresponding random slopes, had very similar coefficients and z-values for all the significant predictors.

Significant main effects (at $p < 0.05$) were found for all but three of the predictors tested: (i) whether the post has been "liked" by another user; (ii) whether the token contains a non-standard spelling; and (iii) the number of phonological segments in the token. Their non-significance is not entirely surprising: (i) KS users seem to "like" posts because of their content, not because of grammatical properties (such as a writer's use of *PtoV*) of which readers may not be consciously aware. (ii) Tokens that were marked as containing a non-standard spelling also included typographical errors, which should have no direct relation to a writer's use of grammatical features. Finally, (iii) although some writers hypothesized that *PtoV* might be disfavored by a general orthographic preference against very long words, the effect for the length of the particle verb (PV) combination, if any, is rather weak.

5.1.1. Effects and Interpretations of Significant Continuous Predictors

Since all continuous predictors were standardized (see their raw distributions in **Figure 4**), the estimates listed in **Table 3** should be interpreted as follows: for every change of one standard deviation of a given effect, the log odds of the *toPV* variant increases (or decreases) by the estimate listed. Visualizations of the predicted effects are provided in **Figure 5**, showing how each of the significant continuous predictors relates to the predicted probability of *toPV*. For each subplot, the predicted probability of *toPV* is plotted at the average level of the other predictors in the model.

One of the more pronounced fixed effects is the number of days that have elapsed since the launch of KS: the more time that has passed (i.e., the more recent the post), the more likely the *toPV* variant is to be used. However, the number of days that have elapsed since user registration (i.e., the user's seniority as a KS member) has an overall *disfavoring* effect on the *toPV* variant. If *toPV* is being used relatively more often over time, then it seems paradoxical for writers to disfavor that variant the longer they interact on the forum. An in-depth discussion of these seemingly contradictory time effects is presented in section 5.2.

The other significant continuous fixed effects are particle entropy and the log frequency of the particle verb combination, which both pattern in ways consistent with the hypotheses outlined above. Particles with higher entropy disfavor the use of *toPV*, suggesting that particles that can more freely associate with different verbs (i.e., more productive particles) are also more tolerant of intervening *tsu* (*PtoV*). More frequent particle verb combinations favor the *PtoV* variant, which was expected under the assumption that high frequency combinations may have a more robust representation of the *PtoV* exemplar in episodic

TABLE 3 | Estimates for fixed effects from logistic regression model of variable order in non-finite particle verbs (n = 37,858), where positive estimates favor the *toPV* variant; significance codes: *** = < 0.001, ** = < 0.01, * = < 0.05, . = < 0.1.

	Estimate	Std. error	*z*-value	*p*-value		*N*
(Intercept)	−2.04	0.10	−19.84	< 0.001	***	37,858
Particle type (vs. other)						10,496
cognate	0.60	0.11	5.72	< 0.001	***	16,307
directional	−0.52	0.11	−4.79	< 0.001	***	11,055
Verb is English borrowing (vs. no)						37,401
yes	0.54	0.19	2.88	0.004	**	457
Post has been “liked” (vs. no)						13,146
yes	−0.07	0.05	−1.63	0.104		24,712
Contains non-standard spelling (vs. no)						30,988
yes	−0.05	0.07	−0.77	0.444		6,870
Persistence (prev. token in post) (vs. none)						26,622
PtoV	−0.53	0.06	−9.15	< 0.001	***	9,749
toPV	0.61	0.07	8.27	< 0.001	***	1,487
Particle entropy (scaled)	−0.33	0.04	−7.90	< 0.001	***	37,858
Num. segments in particle verb (scaled)	0.07	0.04	1.84	0.066	.	37,858
Log frequency of particle verb (scaled)	−0.10	0.03	−3.11	0.002	**	37,858
Days since user registration (scaled)	−0.13	0.06	−2.20	0.028	*	37,858
Days since KS launch (scaled)	0.28	0.06	4.80	< 0.001	***	37,858

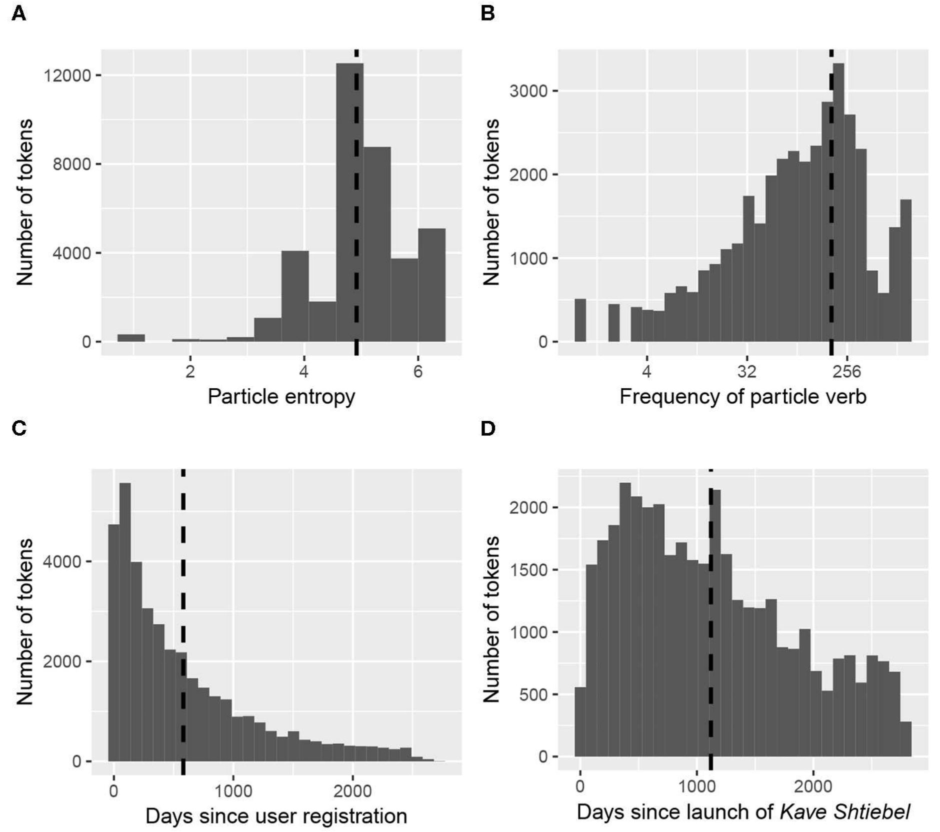


FIGURE 4 | Raw distribution of particle verb tokens across significant continuous predictors (dashed lines represent the means; note that the x-axis of subplot B is on a logarithmic scale) **(A)** Particle entropy. **(B)** Frequency of particle verb. **(C)** Days since user registration. **(D)** Days since launch of *Kave Shtiebel*.

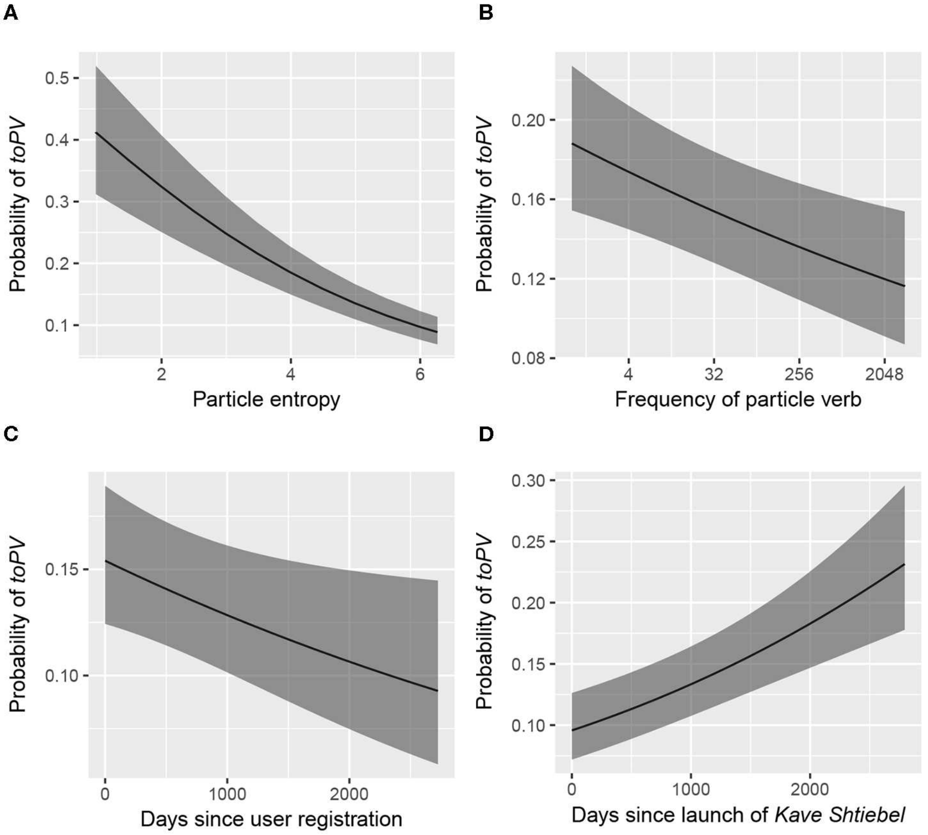


FIGURE 5 | Predicted probability of *toPV* for significant continuous fixed effects (note that the x-axis of subplot B is on a logarithmic scale). **(A)** Particle entropy. **(B)** Frequency of particle verb. **(C)** Days since user registration. **(D)** Days since launch of *Kave Shtiebel*.

memory. Further investigation is needed in order to obtain a clearer picture of the role of frequency in constraining syntactic variation, in Yiddish and in other languages.

5.1.2. Effects and Interpretations of Significant Categorical Predictors

The remaining significant fixed effects (particle type, whether the verb is an English borrowing, and variant persistence) are categorical variables. Their distributions are shown in **Figure 6**. **Figure 7** plots the predicted marginal means, showing how each of the factor levels relates to the predicted probability of *toPV*. Again, for each factor, the predicted probability of *toPV* is plotted at the average level of the other predictors in the model.

Each of these categorical predictors has an effect on the variation in the direction hypothesized. Directional particles, which tend to contribute to the meaning of particle verb combinations in transparent or semantically compositional ways, tolerate the intervention of *tsu* (*PtoV*) at the highest rate. Cognate particles, which are often found in idiomatic or semantically non-compositional combinations, tolerate the intervention of *tsu* at the lowest rate (*toPV*). The "other" particles have an effect that is intermediate between the two types, and significantly different from both. There is a clear effect of whether the verb is an English borrowing, such that borrowed verbs favor *toPV* relative to other kinds of verbs. Note, however, that there is a massive imbalance across borrowings and non-borrowings (see **Figure 6B**), and consequently this effect should be interpreted with some caution. For example, for certain tokens tagged as having "English verbs," it is actually the entire particle verb combination that is a borrowing, and in English the "particle" is actually an inseparable prefix (e.g., *op-deyt-n* 'update'; cf. *date up). These tokens understandably favor *toPV* (though never at 100%; e.g., there are 8 tokens of the *PtoV* variant *op-tsu-deyt-n* compared to 40 tokens of *tsu op-deyt-n*). Finally, there is a clear effect of persistence from the variant most recently used in the post, such that users are biased to repeat the same variant whether *PtoV* or *toPV*. Tokens of "none" are situated in the middle. This is to be expected, both because the absence of a previous token should not give rise to any persistence effect, and because the data are distributed in such a way that the majority of tokens are the first (or only token) of their respective posts (see **Figure 6C**). These findings lend themselves to follow-up analysis considering whether texts written for distribution on the internet (in Yiddish or any other language) generally exhibit stronger persistence effects than other genres of audience-oriented writing, in which the effects of cognitive constraints on variation may be tempered by more careful editing.

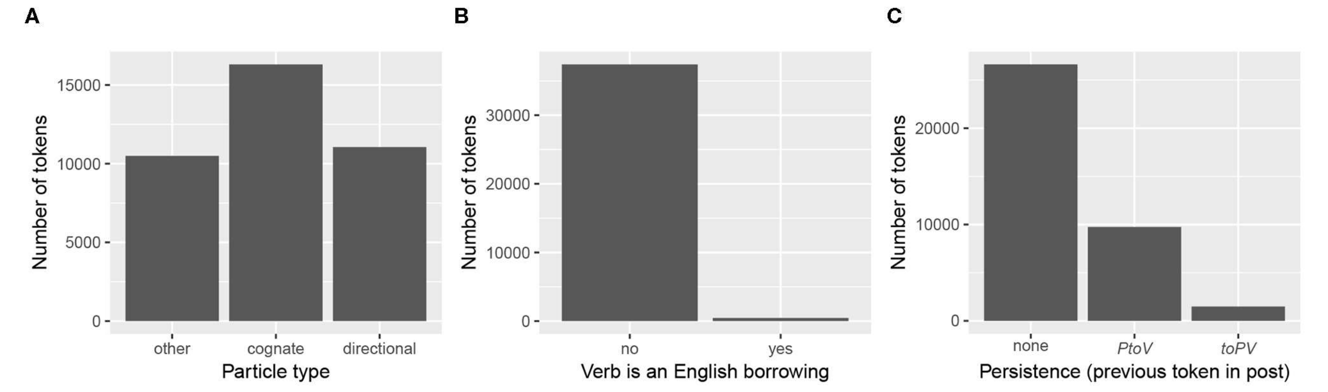


FIGURE 6 | Raw distribution of particle verb tokens across significant categorical predictors. **(A)** Particle type. **(B)** Verb is an English borrowing. **(C)** Persistence (previous token in post).

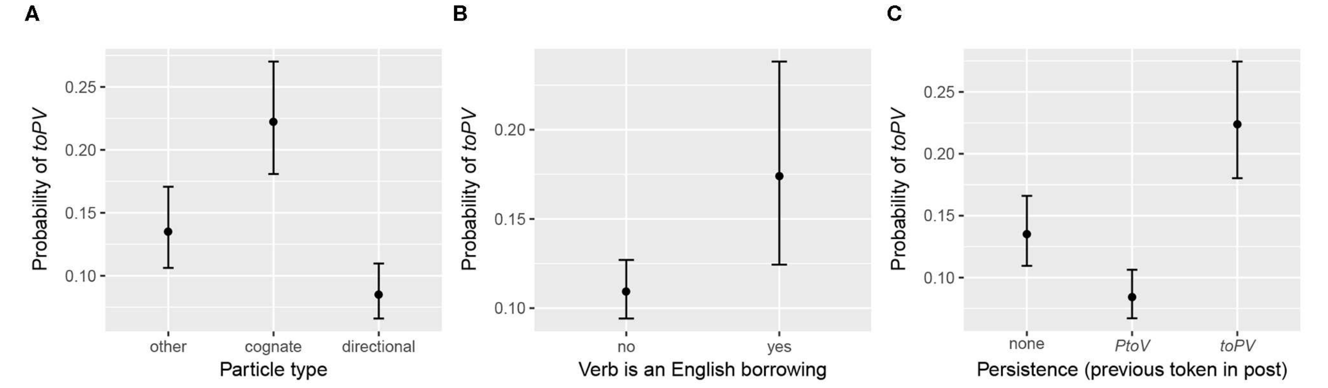


FIGURE 7 | Predicted probability of *toPV* for significant categorical fixed effects. **(A)** Particle type. **(B)** Verb is an English borrowing. **(C)** Persistence (previous token in post).

5.2. Discussion of Syntactic Change in Real Time

To reiterate one of the more intriguing findings of the statistical analysis, a seemingly contradictory effect was identified for the time elapsed since user registration and for the time elapsed since the launch of KS: users favor the standard *PtoV* variant the older their accounts are, despite a forum-wide trend favoring the non-standard *toPV* variant in real time. In other words, there seems to be evidence both for *individual change* toward greater use of *PtoV* and *community change* toward greater use of *toPV*.

5.2.1. Implicit Standardization Favoring *PtoV* in Real Time

The finding that increased user seniority favors *PtoV* is consistent with the observation that online platforms, and KS in particular, have created new opportunities for Hasidic men to acquire experience and skill as Yiddish writers. In a sociolinguistic interview, one KS user Fayvl (31; Williamsburg) explicitly connected the advent of discussion forums to the proliferation of written standards:

> *Kave Shtiebel* is trying to… the leaders of it, I don't know who they are, are trying to make Yiddish a, that it should have rules… It has changed quite a lot, actually. Because when I grew up, I mean, before the internet, there wasn't anywhere to write in Yiddish. A Hasid who wanted to write, he didn't have anywhere to write. You understand? Because... there just wasn't [any outlet]. Today you can write on the internet, or WhatsApp. We want to be able to write well. Automatically it's becoming a language, you know? The language is being formed from scratch, in a certain sense. (Translated from Yiddish.)

Although the mention of "rules" here encompasses norms of spelling, punctuation, and vocabulary, Fayvl's view also offers a cogent explanation for the empirical finding that more experienced writers favor a conservative variant in syntax. The longer users spend on KS posting messages and interacting with other KS writers, the likelier it is that they will acquire the norms used by others, including grammatical norms.

One of the distinct advantages of using a discussion forum as a linguistic corpus is that every post has a timestamp and every user has a registration date. This makes it trivial to organize users into cohorts and track their behavior over time—akin to a longitudinal panel study of spoken language across age cohorts. The approach pursued here is to group users based on year of account registration. Because the number of new KS users has stabilized since the forum's launch in 2012 (**Figure 8**), we collapse the most recent years (2015–now) into a single cohort.

Figure 9 shows that for the largest single-year cohorts (2012, 2013, and 2014), who produced 81.7% of all tokens of non-finite particle verbs, users enter the forum with an increasingly high rate of *toPV*, which then falls over time. This suggests that regardless of when a cohort joins the forum, and regardless of what their initial rate of *toPV* is, by virtue of interacting with other users they seem to be acquiring the norm that associates *PtoV* with standard or "correct" usage. (The cohort since 2015 shows an increase in *toPV*, but the trend is flatter overall; if norms are being acquired implicitly, perhaps more time is required to see a decrease.)

Unlike inconsistencies in spelling, which are the object of explicit commentary online and offline, syntactic variation tends to fly under the radar of most writers. To my knowledge, there has been no discussion of the variation between *PtoV* and *toPV* on KS or any other Hasidic discussion forum. For this reason, and because the trend is observable even within single-year user cohorts, I take the finding about user seniority as empirical evidence of *rapid implicit standardization* among KS users.

If standardization is taking place on Hasidic social media more generally, the effect may actually be amplified on KS, where a writer's adherence to norms in spelling and punctuation is viewed as a sign that he is mature, intellectual, and worldly. These are qualities that are especially valued on KS, a forum that positions itself as challenging the Hasidic mainstream, particularly the perception of Hasidic "groupthink" which is so often criticized on the forum. Additional research using data from other forums could shed light on the factors motivating implicit standardization among Hasidic Yiddish writers.

5.2.2. Community Change Favoring *toPV* in Real Time

If users favor the standard *PtoV* variant the longer their accounts remain open and active, it seems strange that there should also be a real-time effect favoring non-standard *toPV* on the forum overall. While it is possible that we are witnessing a genuine change in progress, one that reflects a possible increase in *toPV* in spoken Yiddish, it is surprising to find such an effect on a forum that has existed for under eight years, and whose users may not differ in age even if they joined the site at different times.

The contradiction is resolved if we acknowledge that there may be significant differences in the social characteristics of users depending on how recently they began writing on KS. As **Figure 8** shows, a large number of users registered on KS within the first month or so of its launch. Because KS was founded as an offshoot of a different forum, *iVelt*, most of these early users already had a history of communicating in written Yiddish—certainly on *iVelt* if not on other online platforms, too. It stands to reason that these early users may have had a lower initial rate of *toPV* when KS first launched, since their development as Yiddish writers actually began elsewhere. (This is supported in **Figure 9** by comparing the initial probability of *toPV* in the 2012 cohort against the subsequent cohorts from 2013 and 2014). If this view is correct, then a 36-year-old Hasidic Jew who registers on KS for the first time in 2019 may be much less experienced than a 36-year-old who joined KS seven years earlier. This could account for the conflicting trends in real-time data, where newcomers to the forum favor *toPV* even though individual users are expected to favor *PtoV* as they gain experience and facility with the norms of written Yiddish. Impressionistically, this explanation is supported by the fact that newcomers' welcome messages to the subforum *lomikh zikh forshteln far aykh* 'let me introduce myself to you' are substantially less standard in orthography and vocabulary than one finds among more senior writers. To test this explanation more directly, a follow-up study could compare the "standardness" of written Yiddish across different seniority levels on KS, in terms of users' grammatical norms as well as orthography and vocabulary.

6. CONCLUSIONS

While sociolinguists have acknowledged the hegemony of English in quantitative studies of variation, work on minority language varieties is still underrepresented (Meyerhoff and Nagy, 2008; Stanford, 2016; Guy and Adli, 2019). The shortage of research on these languages is especially pronounced in areas of linguistics where new computational methods have made it possible to identify complex trends in large messy datasets. As Nicholas Ostler has argued, "just as [the Yiddish philologist] Max Weinreich once remarked that a language is a dialect with an army and a navy, nowadays a language is a dialect with a dictionary, grammar, parser, and a multi-million-word corpus of texts, which are computer tractable, and ideally a speech database too" (Ostler, 2011, p. 320). As these computational resources continue to be developed in Hasidic Yiddish and other minority language varieties, corpus research will be able to uncover significant linguistic and social constraints on variability in a larger number of the world's languages.

This analysis of syntactic variation on a Hasidic Yiddish discussion forum has revealed that the choice of the *PtoV* or *toPV* order in non-finite particle verbs—seemingly arbitrary, given the presence of near-minimal pairs with equivalent semantics—is conditioned by both linguistic and social factors. The conditioning effects are also consistent with the findings from studies of particle verb variation in English. For example, the statistical analysis identified significant effects for particle type, which is taken to approximate the degree of semantic transparency, and for particle entropy, which is taken to approximate particle productivity across different verbs. Additional comparative studies are needed if variationists seek to evaluate the cross-linguistic applicability of conditioning factors assumed to be universal, e.g., the tendency to minimize syntactic and semantic dependencies (Lohse et al., 2004) or the tendency to repeat recent variants (Tamminga, 2016).

That some of the factors influencing particle verb variation in English also play a role in Hasidic Yiddish begs the question: Are these overlapping constraints due to universal linguistic properties, or is it possible that they arose in Yiddish due to contact with English? The latter hypothesis is consistent with an

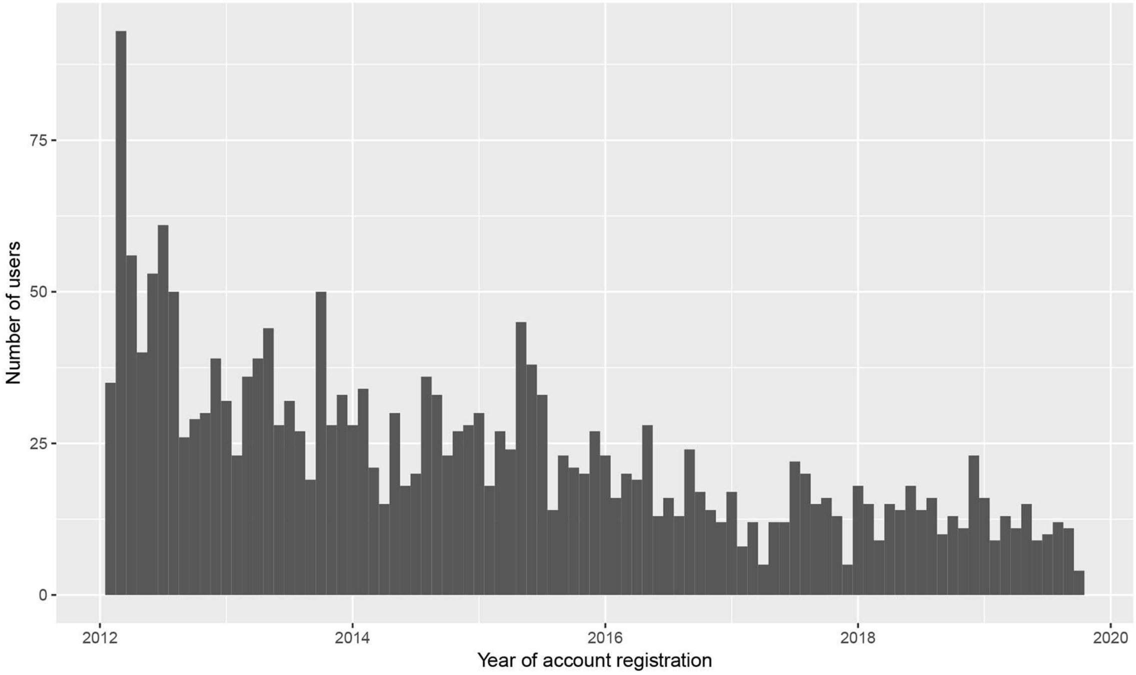


FIGURE 8 | Users of *Kave Shtiebel* according to date of account registration (binned by month).

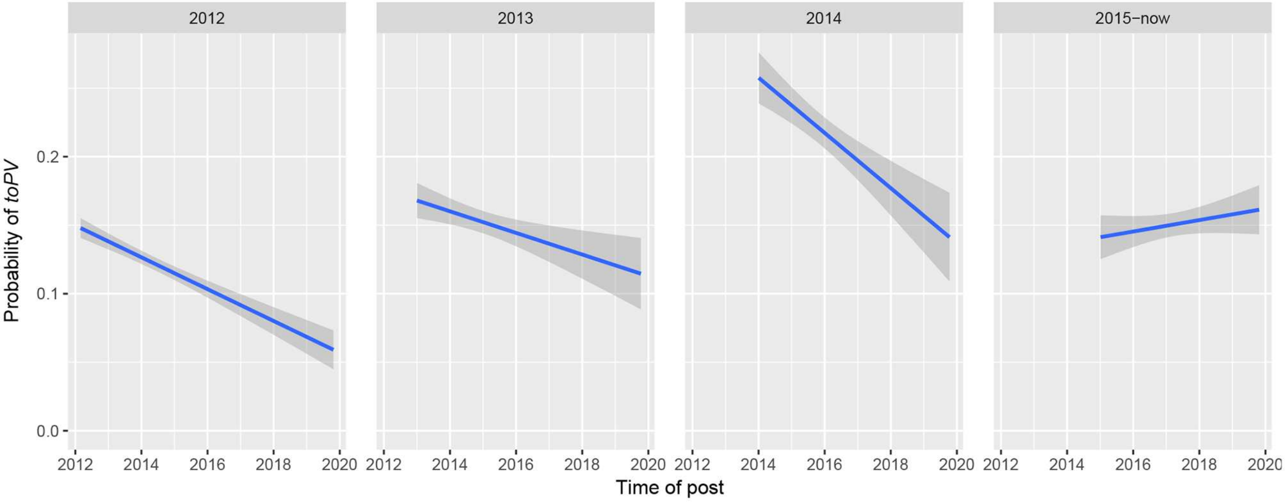


FIGURE 9 | Regression lines showing the changing probability of *toPV*, based on plots of the raw distribution of tokens over time; data separated by the calendar year in which user registered on KS.

assumption widely held by Yiddish scholars and speakers alike, that *all* changes taking place in American Yiddish must ultimately derive from contact with English. In fact, some of the Hasidic men consulted during this project assured me that *toPV* is itself a structural borrowing from English, since *to* always comes before the verb in English. However, this explanation ignores the fact that *tsu* 'to' always precedes the verb in Yiddish as well, as shown in (3) for infinitives without particles.

In the absence of compelling evidence corroborating the English contact-based model, I maintain that the increased probability of *toPV* could be a Yiddish-internal development. First, although relatively rare, tokens of *toPV* can be found in pre-Holocaust Yiddish publications from Eastern Europe. In fact, some of the earliest examples of *toPV* come from traditional glosses of religious texts in Hebrew (Simon Neuberg, pers. comm.), such as Rashi's commentary on Genesis 14:9

*mi**lirdoyf** akhareyhem* 'from chasing after them,' glossed in Yiddish as *fun **tsu nokh yogn** zey* (lit., from **to after chase** them)[17]. Traditional Hebrew glossing, also known as *kheyder-taytsh* 'school translation,' often preserves the morpheme or word order of the Hebrew even if the resulting Yiddish is somewhat awkward structurally. The influence of such glosses on the development of Yiddish has been posited before (Timm, 2005), and it is plausible that the *l*-prefix marking Hebrew infinitives played some role in the emergence of *toPV*. The effect might be especially pronounced among Orthodox Jewish men, who were—and still are—exposed to such glosses in their *kheyder* education.

Second, separable particles never appear preverbally in English (*toVP*: *to throw up*; cf. *toPV *to up throw* and *PtoV *up to throw*), whereas particles invariably precede the verb in Yiddish infinitives. Third, the variation in English involves the relative ordering of particles and full noun phrase objects, and it is not limited to non-finite contexts (*I will call {up} the mayor {up}*; *I called {up} the mayor {up}*, etc.). In Yiddish, however, the relative ordering of particles and full noun phrase objects is generally fixed in the present tense, when verb-second (V2) movement causes the particle to appear postverbally:

(10) a. er **ruft on** dem melamed.
he **calls on** the teacher
'He is calling up the teacher (on the phone).'
b. ?er **ruft** dem melamed **on**.
he **calls** the teacher **on**
'He is calling the teacher up (on the phone).'

It is conceivable that Yiddish borrowed some of the underlying constraints on particle verb variation from English without borrowing its variant surface structures. However, it seems more plausible that the overlap in conditioning factors stems from language-independent considerations, which can be posited for all of the (non-social) predictors selected in the statistical model.

With respect to socio-stylistic constraints, the analysis revealed that a single online discussion forum can be a vehicle both for the spread of an innovative linguistic form and for the reinforcement of conservative written standards. This finding contributes to our understanding of the role that social media sites play in the rapid diffusion of linguistic change (e.g., Eisenstein et al., 2014). Given popular stereotypes about the internet as a place where language is "ruined"—where non-standard abbreviations, acronyms, and slang are spread—it is surprising that a discussion forum could be a venue for the proliferation of written norms. Perhaps *implicit* standardization is only possible in a language community that does not have a formal system for teaching and enforcing such written norms. Alternatively, implicit standardization could be a more general phenomenon affecting online writing, but researchers' focus on short-form media (such as text messages and tweets) has obscured this fact. Large corpus studies, especially of other minority language varieties, could shed light on this question of how language change occurs online, whether that change involves an increase or a decrease in the use of standard variants.

Finally, this study has demonstrated that robust patterns of language variation and change can be gleaned from a relatively modest online community of writers, using data drawn from posts written over a period of less than eight years. Even if the challenge of data scarcity looms large for machine translation in "low-resource" minority languages (Genzel et al., 2009)[18], it should not deter sociolinguists from attempting to analyze variation in those languages. This result should inspire confidence that corpus sociolinguistics can uncover patterns of grammatical variation and change in minority language varieties, provided that specialists know where to find raw data and can define heuristics to identify tokens of variables. Studies of variation on social media platforms not only elucidate linguistic behavior on the internet, but they also generate testable hypotheses for research conducted in the speech community.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by New York University, University Committee on Activities Involving Human Subjects. All participants provided their written informed consent.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the National Science Foundation under Grant No. DGE 1342536. Publication was made possible in part by support from the Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

ACKNOWLEDGMENTS

I would like to acknowledge Sarah Bunin Benor, Gregory R. Guy, Laurel MacKenzie, Simon Neuberg, Chaya R. Nove, Lea Schäfer, Scott Seyfarth, and audiences at New Ways of Analyzing Variation 46, the 2018 meeting of the Linguistic Society of America, the Sociolinguistics Lab at NYU, and the Sociolinguistics Lab at Berkeley for helpful comments and suggestions. I would also like to thank the founders of *Kave Shtiebel* for granting me access to their database of public posts.

[17] This particular example was found in a Yiddish edition of the *mikroes-gdoyles* (the Hebrew Bible along with various commentaries) published in Vilnius in 1899 but which may be based on an older translation. Available online: https://books.google.com/books?id=7W4_AAAAYAAJ, p. 122.

[18] Thanks to improvements in optical character recognition, the Yiddish Book Center now supports text search across its collection of over 10,000 volumes (ocr.yiddishbookcenter.org). This could be a boon for machine translation in Yiddish.

REFERENCES

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *J. Sociolinguist*. 18, 135–160. doi: 10.1111/josl.12080

Bannard, C. (2005). Learning about the meaning of verb-particle constructions from corpora. *Comput. Speech Lang*. 19, 467–478. doi: 10.1016/j.csl.2005.02.003

Baroni, M., and Bernardini, S. (2004). "BootCaT: bootstrapping corpora and terms from the web," in *Proceedings of the Language Resources and Evaluation Conference (LREC) 2004* (Lisbon), 1313–1316.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1–48. doi: 10.18637/jss.v067.i01

Bayley, R., Greer, K., and Holland, C. (2013). Lexical frequency and syntactic variation: a test of a linguistic hypothesis. *Univers. Pennsylvania Working Pap. Linguist*. 19, 21–30.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.

Biskup, P., Putnam, M., and Smith, L. C. (2011). German particle and prefix verbs at the syntax-phonology interface. *Leuvense Bijdragen-Leuven Contrib. Linguist. Philol*. 97, 106–135. doi: 10.2143/LB.97.0.2977249

Bleaman, I. L. (2018). *Outcomes of minority language maintenance: variation and change in New York Yiddish* (Ph.D. thesis). New York University, New York, NY, United States.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Lang. Variat. Change* 14, 261–290. doi: 10.1017/S0954394502143018

Conrod, K., Tatman, R., and Koncel-Kedziorski, R. (2016). "We who tweet: pronominal relative clauses on Twitter," in *Proceedings of Corpus Linguistics Fest (CLiF) 2016*, eds S. Kübler and M. Dickinson (Bloomington, IN), 17–21.

Dehé, N. (2002). *Particle Verbs in English: Syntax, Information Structure and Intonation*. Amsterdam; Philadelphia, PA: John Benjamins. doi: 10.1075/la.59

Dehé, N. (2015). "Particle verbs in Germanic," in *Word-Formation: An International Handbook of the Languages of Europe*, Vol. 1, eds P. O. Müller, I. Ohnheiser, S. Olsen, and F. Rainer (Berlin: De Gruyter Mouton), 611–626.

Deutsch, N. (2009). The forbidden fork, the cell phone Holocaust, and other Haredi encounters with technology. *Contemp. Jewry* 29, 3–19. doi: 10.1007/s12397-008-9002-7

Diesing, M. (1997). Yiddish VP order and the typology of object movement in Germanic. *Nat. Lang. Linguist. Theory* 15, 369–427. doi: 10.1023/A:1005778326537

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *J. Sociolinguist*. 19, 161–188. doi: 10.1111/josl.12119

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE* 9:e113114. doi: 10.1371/journal.pone.0113114

Erker, D., and Guy, G. R. (2012). The role of lexical frequency in syntactic variability: variable subject personal pronoun expression in Spanish. *Language* 88, 526–557. doi: 10.1353/lan.2012.0050

Fader, A. (2009). *Mitzvah Girls: Bringing Up the Next Generation of Hasidic Jews in Brooklyn*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400830992

Fader, A. (2017). Ultra-Orthodox Jewish interiority, the Internet, and the crisis of faith. *HAU J. Ethnogr. Theory* 7, 185–206. doi: 10.14318/hau7.1.016

Genzel, D., Macherey, K., and Uszkoreit, J. (2009). "Creating a high-quality machine translation system for a low-resource language: Yiddish," in *Machine Translation Summit XII*, ed L. Gerber (Ottawa, ON: Association for Machine Translation in the Americas), 1–8.

Gold, E. (1999). *Aspect, tense and the lexicon: expression of time in Yiddish* (Ph.D. thesis). University of Toronto, Toronto, ON, Canada.

Gries, S. T. (2001). A multifactorial analysis of syntactic variation: particle movement revisited. *J. Quant. Linguist*. 8, 33–50. doi: 10.1076/jqul.8.1.33.4092

Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *J. Psycholinguist. Res*. 34, 365–399. doi: 10.1007/s10936-005-6139-3

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on American social media. *J. English Linguist*. 46, 293–319. doi: 10.1177/0075424218793191

Guy, G. R., and Adli, A. (2019). A manifesto on cross-cultural sociolinguistics: The Fourth Wave in the study of language variation and change. Unpublished manuscript.

Haddican, B., and Johnson, D. E. (2012). Effects on the particle verb alternation across English dialects. *Univers. Pennsylvania Working Pap. Linguist*. 18, 31–40.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Comput. Environ. Urban Syst*. 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Ilbury, C. (2019). "Sassy Queens": stylistic orthographic variation in Twitter and the enregisterment of AAVE. *J. Sociolinguist*. 24, 245–264. doi: 10.1111/josl.12366

Isaacs, M. (1999). Haredi, *haymish* and *frim*: Yiddish vitality and language choice in a transnational, multilingual community. *Int. J. Sociol. Lang*. 138, 9–30. doi: 10.1515/ijsl.1999.138.9

Jacobs, N. G. (2005). *Yiddish: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Jones, R. J., Cunliffe, D., and Honeycutt, Z. R. (2013). Twitter and the Welsh language. *J. Multiling. Multicult. Dev*. 34, 653–671. doi: 10.1080/01434632.2013.812096

Katz, D. (1993). *Tikney takones: Fragn fun yidisher stilistik [Amended Amendments: Issues in Yiddish Stylistics*]. Oxford: Oksforder Yidish.

Keegan, T. T., Mato, P., and Ruru, S. (2015). Using Twitter in an indigenous language: an analysis of te reo Māori tweets. *AlterNative Int. J. Indig. Peoples* 11, 59–75. doi: 10.1177/117718011501100105

Kroch, A., and Small, C. (1978). "Grammatical ideology and its effect on speech," in *Linguistic Variation: Models and Methods*, ed D. Sankoff (New York, NY: Academic Press), 45–55.

Lohse, B., Hawkins, J. A., and Wasow, T. (2004). Domain minimization in English verb-particle constructions. *Language* 80, 238–261. doi: 10.1353/lan.2004.0089

Mark, Y. (1978). *Gramatik fun der yidisher klal-shprakh [A Grammar of Standard Yiddish*]. New York, NY: Congress for Jewish Culture.

Meyerhoff, M., and Nagy, N. (2008). "Introduction: social lives in language," in *Social Lives in Language—Sociolinguistics and Multilingual Speech Communities: Celebrating the Work of Gillian Sankoff*, eds M. Meyerhoff and N. Nagy (Amsterdam; Philadelphia, PA: John Benjamins), 1–16. doi: 10.1075/impact.24.02nag

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., and Vespignani, A. (2013). The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* 8:e61981. doi: 10.1371/journal.pone.0061981

Nguyen, D., Trieschnigg, D., and Cornips, L. (2015). "Audience and the use of minority languages on Twitter," in *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (Palo Alto, CA), 666–669.

Nove, C. R. (2018). The erasure of Hasidic Yiddish from twentieth century Yiddish linguistics. *J. Jewish Lang*. 6, 111–143. doi: 10.1163/22134638-06011142

Ostler, N. (2011). "Language maintenance, shift, and endangerment," in *The Cambridge Handbook of Sociolinguistics*, ed R. Mesthrie (Cambridge: Cambridge University Press), 315–334. doi: 10.1017/CBO9780511997068.024

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cogn. Linguist*. 7, 149–182. doi: 10.1515/cogl.1996.7.2.149

Schaechter, M. (1995). *Yidish tsvey: A lernbukh far mitndike un vaythalters [Yiddish II: An Intermediate and Advanced Textbook*]. New York, NY: League for Yiddish.

Scherre, M. M. P. (2001). Phrase-level parallelism effect on noun phrase number agreement. *Lang. Variat. Change* 13, 91–107. doi: 10.1017/S0954394501131042

Schnoebelen, T. (2008). "Measuring compositionality in phrasal verbs," in *Third Workshop on Quantitative Investigations in Theoretical Linguistics (QITL3)*, eds A. Arppe, K. Sinnemäki, and U. Nikanne (Helsinki), 58–61.

Silverstein, M. (1981). *The Limits of Awareness (Sociolinguistic Working Paper Number 84)*. Austin, TX: Southwest Educational Development Laboratory.

Stanford, J. N. (2016). A call for more diverse sources of data: variationist approaches in non-English contexts. *J. Sociolinguist*. 20, 525–541. doi: 10.1111/josl.12190

Talmy, L. (2000). *Toward a Cognitive Semantics, Vol. 2: Typology and Process in Concept Structuring*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/6848.001.0001

Tamminga, M. (2016). Persistence in phonological and morphological variation. *Lang. Variat. Change* 28, 335–356. doi: 10.1017/S0954394516000119

Tamminga, M., MacKenzie, L., and Embick, D. (2016). The dynamics of variation in individuals. *Linguist. Variat*. 16, 300–336. doi: 10.1075/lv.16.2.06tam

Timm, E. (2005). *Historische jiddische Semantik: Die Bibelübersetzungssprache als Faktor der Auseinanderentwicklung des jiddischen und des deutschen Wortschatzes [Historical Yiddish Semantics: The Bible Translation Language as a Factor in the Divergence of Yiddish and German Vocabulary*]. Tübingen: Max Niemeyer. doi: 10.1515/9783110945034

U.S. Census Bureau (2015). *2015 American Community Survey, B16001: Language Spoken at Home by Ability to Speak English for the Population 5 Years and Over*.

Weiner, E. J., and Labov, W. (1983). Constraints on the agentless passive. *J. Linguist*. 19, 29–58. doi: 10.1017/S0022226700007441

Yates, S. J. (1996). "Oral and written linguistic aspects of computer conferencing: a corpus based study," in *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, ed S. C. Herring (Amsterdam; Philadelphia, PA: John Benjamins), 29–46. doi: 10.1075/pbns.39.05yat

YIVO (Yidisher visnshaftlekher institut) (1999). *Der eynheytlekher yidisher oysleyg [The Standardized Yiddish Orthography*]. YIVO Institute for Jewish Research and the League for Yiddish, New York, NY.

Zeller, J. (2001). *Particle Verbs and Local Domains*. Amsterdam; Philadelphia, PA: John Benjamins.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 02 July 2020
doi: 10.3389/frai.2020.00046

Size Matters: Digital Social Networks and Language Change

Mikko Laitinen[1,2]*, *Masoud Fatemi*[1,2] *and Jonas Lundberg*[2]

[1] *School of Humanities/English, University of Eastern Finland, Kuopio/Joensuu, Finland,* [2] *Center for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden*

Social networks play a role in language variation and change, and the social network theory has offered a powerful tool in modeling innovation diffusion. Networks are characterized by ties of varying strength which influence how novel information is accessed. It is widely held that weak-ties promote change, whereas strong ties lead to norm-enforcing communities that resist change. However, the model is primarily suited to investigate small ego networks, and its predictive power remains to be tested in large digital networks of mobile individuals. This article revisits the social network model in sociolinguistics and investigates network size as a crucial component in the theory. We specifically concentrate on whether the distinction between weak and strong ties levels in large networks over 100 nodes. The article presents two computational methods that can handle large and messy social media data and render them usable for analyzing networks, thus expanding the empirical and methodological basis from small-scale ethnographic observations. The first method aims to uncover broad quantitative patterns in data and utilizes a cohort-based approach to network size. The second is an algorithm-based approach that uses mutual interaction parameters on Twitter. Our results gained from both methods suggest that network size plays a role, and that the distinction between weak ties and slightly stronger ties levels out once the network size grows beyond roughly 120 nodes. This finding is closely similar to the findings in other fields of the study of social networks and calls for new research avenues in computational sociolinguistics.

Keywords: social networks, Twitter, bot exclusion, data mining, weak ties, social network size

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Annibale Elia,
University of Salerno, Italy
Xavier Blanco Escoda,
Autonomous University of
Barcelona, Spain

****Correspondence:***
Mikko Laitinen
mikko.laitinen@uef.fi

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *09 March 2020*
Accepted: *27 May 2020*
Published: *02 July 2020*

Citation:
Laitinen M, Fatemi M and Lundberg J
(2020) Size Matters: Digital Social
Networks and Language Change.
Front. Artif. Intell. 3:46.
doi: 10.3389/frai.2020.00046

INTRODUCTION

This article focuses on social networks and explores network size as a key determinant in the network theory used in sociolinguistics. Building on Granovetter (1973), the theory postulates that individuals form personal communities that provide a meaningful framework for them in their daily life (Milroy and Llamas, 2013). An individual's social network is the sum of relationships contracted with others, and a network may be characterized by ties of varying strength. If ties are strong and multiplex, the network is dense, and individuals are linked through close ties (such as friends). Conversely, ties can be weak in which case individuals are predominantly linked through occasional and insignificant ties (such as acquaintances), and the network is loosely knit. Most importantly, networks contribute to language maintenance and change. Ample empirical evidence shows that loose-knit networks promote innovation diffusion, whereas dense multiplex networks lead to communities that resist change (Milroy and Milroy, 1978, 1985; Milroy, 1987; Lippi-Green, 1989). The underlying reason for the weakness of strong ties in transmitting innovation is the fear of losing one's social standing in a network. Adopting new ideas is socially risky, and we do not want to "rock the boat" in dense social structures.

Even though the social network theory is influential in sociolinguistics, it is mostly based on small data. Most studies have focused on what are usually referred to as ego networks obtained using ethnographic observation. According to Milroy and Milroy (1992, p. 5) this "effectively limits the field of study, generally to something between 30 and 50 individuals." Moreover, it has been suggested that the quantitative variable of a network "cannot be easily operationalized in situations where the population is socially and/or geographically mobile" (Milroy, 1992, p. 177). In this paper, we concentrate on networks that are larger than small networks of only a few dozen of individuals. This has been done because evidence from social anthropology suggests that average human networks are substantially larger, and individuals can maintain networks with well over 200 nodes (McCarty et al., 2001). Prior empirical work in sociolinguistics has therefore covered only a limited section of possible network sizes.

We have two research foci. First, we test the extent to which social media data from Twitter and computational methods could be utilized to operationalize network ties of highly mobile individuals in very large datasets. Second, we specifically concentrate on the effect of network size on the validity of the theory. We investigate if the fear of losing one's social standing by "rocking the boat" disappears in large strong-tie networks.

To respond to these questions, we discuss two computational methods that can take up large and messy social media data and render them usable for analyzing networks in sociolinguistics, thus expanding the empirical basis from small-scale ethnographic observations. The first method aims at uncovering broad quantitative patterns in data and utilizes what we call a cohort-based method of network size. The second consists of an algorithm-based approach that uses mutual interaction parameters in Twitter and aims to verify the patterns obtained using the cohort-based approach.

By doing so, the article continues our pilot investigation, which suggests that network size is a crucial component in the theory. Our first results indicated that weak ties are meaningful in small networks, but the distinction between truly weak ties and slightly stronger ties levels out when network size increases beyond a certain threshold level (Laitinen et al., 2017). This pilot was based on social media data that had not yet been cleaned of unwanted software robot data (i.e., bots). In the present study, we attempt to replicate the study using a more accurate dataset from which we have removed bots by means of machine-learning techniques and by using novel computational methods to test our first observations. Bot content can result in inaccuracies, and previous computational sociolinguistic studies rely on a range of methods when bots are handled. Their presence may be recognized, but they are nevertheless included in the results (Huang et al., 2016; Laitinen et al., 2017). Other methods, such as excluding material by using metadata parameters, are occasionally used (Coats, 2017), but as we demonstrate below in section Material and Methods, more advanced solutions are available.

As shown in the next section, the role of network size in sociolinguistics is an understudied phenomenon, which not only requires new tools but could also shed light on the contrast between strong and weak ties in innovation diffusion. One example is that while the weak-tie model is beneficial, it has recently seen substantial theoretical elaboration, and recent advances have broadened the understanding of networks ties as a unidimensional concept (Aral and Van Alstyne, 2011). What is clear is that weak-tie and close-knit networks are different for small ego networks obtained through ethnographic methods, but if network size is ignored, the social network theory is not fully consistent with some of the major findings in sociolinguistics. First, it is widely held that there is one period when individuals maintain maximally close ties with their peers, and that is adolescence (Chambers, 2003, p. 90–91). Yet, the role of adolescents in language change is indisputable and verified in both real-time and apparent-time studies of change in progress (Labov, 2001, p. 76; Tagliamonte and D'Arcy, 2009). There might, of course, be other reasons than interpersonal ties during adolescence that lead teens to diverge from adult norms, but network size deserves to be studied in more detail. Moreover, ample macro-level evidence suggests that densely populated and sufficiently large working-class urban areas have, throughout history, been sites for innovations (e.g., the Jewish quarters all over Europe, Harlem in New York City, or St. John's Ward in Toronto). Pan et al. (2013) suggest that it is the size and density of the ties of a center that are crucial for information diffusion. They investigate social-tie density and information contagion in urban populations, and their quantitative model shows how density, with both weak and strong ties, drives the "super-linear" growth of interaction and information diffusion. Close-knit urban centers may, of course, be sufficiently large to sustain individuals with weak ties through whom innovations spread to a community, but we simply do not yet know whether the role of weak and strong ties levels out beyond a certain threshold level.

Section Social Networks in Variationist Sociolinguistics Discusses not only the theoretical basis of social networks in sociolinguistics but also reviews recent insight from complex systems analysis and social network theory. Section Material and Methods details the material and the two methodologies. Section Results presents the results, and, finally, section Conclusions discusses the implications of our findings.

SOCIAL NETWORKS IN VARIATIONIST SOCIOLINGUISTICS

Social network analysis in the variationist paradigm transpires from the idea that individuals establish interpersonal ties of varying strengths to form communities. These personal social networks are not independent from other socio-cultural frameworks but are closely related to other variables, such as gender and social layer (Milroy and Milroy, 1992). Interpersonal ties influence the rate at which innovations are adopted and how they diffuse into a community. Sociolinguists have shown that strong networks tend to maintain and support local norms and provide resistance to the adoption of competing norms from the outside. Conversely, conditions that are characterized by weak and uniplex ties are important channels for outside influence as people in such situations tend to accommodate to each other linguistically. Contact situations with weak ties therefore contribute positively to the spread of innovations.

This finding builds on Granovetter's (1973, p. 1365) observation that "only weak ties may be local bridges." More people can be reached through weak ties, but not all weak ties serve this function, "only those acting as bridges between network segments" (1983, p. 229). To explain this somewhat counterintuitive observation, Granovetter (1973, 1983) argues that close-knit networks encourage local cohesion and norm-enforcing communities where the adoption of innovations is risky. Loose-knit networks with individuals already on the social fringes are more susceptible to external innovations. In addition, weak ties may be expected to be more numerous among mobile individuals and are thus more likely to contribute to the diffusion of an innovation.

In variationist sociolinguistics, network ties have been operationalized in various ways (Milroy and Llamas, 2013). In the Belfast study, they were measured using five indicators to establish how complex and dense a particular tie was. The indicators consisted of (a) having membership in a locally-based group, (b) having ties with at least two households in the neighborhood, (c) sharing a workplace with two or more individuals from the neighborhood, (d) sharing a workplace with same-sex individuals from the neighborhood, and (e) being involved in voluntary activities with individuals from the same workplace. The responses resulted in a network strength scale, which formed an independent variable, and these values were compared to the dependent (phonological) variables. The results show that the individuals with strong network ties with the local community also exhibited the highest share of local, vernacular speech, and "that a close-knit network has an intrinsic capacity to function as a norm-enforcement mechanism, to the extent that it operates in opposition to larger scale institutional standardizing pressures" (Milroy and Milroy, 1985, p. 359).

A large body of variationist sociolinguistic literature exists in which the network-based approach has been applied to small contemporary communities (Milroy and Llamas, 2013). Milroy and Milroy (1978) use 46 speakers from three urban, blue-collar Belfast communities, and the network ties were established through a participant observation process in which a researcher was introduced to a community by means of a friend-of-a-friend technique. Of these, 12 had network scores qualifying them as weak tie individuals. The same also applies to Granovetter's (1973, p. 1368–1371) study as his empirical data came from a random sample of 100 personal interviews taken from the total sample of 282. Carefully constructed personal networks are obviously important, but the availability of social media data also forces us to ask if the model holds when tested with considerably larger networks.

Network size has not been considered as a separate independent variable in variationist sociolinguistics (Milroy and Llamas, 2013). The model has been applied to large communities in macro-level approaches (Milroy and Milroy, 1985; contrasting Icelandic and English; Raumolin-Brunberg, 1996; investigating mobility as a result of the Civil War in the seventeenth-century England, and Nevalainen, 2000; examining patterns of mobility in Early Modern London). However, while all of these studies are rich in linguistic evidence, they nevertheless contain no direct quantitative evidence of how much weak ties actually increase in the settings that are examined. They rely on indirect evidence of migration patterns, population growth and birth/death rates for instance, but information of average network size per community is not detailed.

Recent findings in social anthropology have shown that an average network size is larger than a few dozen individuals. Dunbar (1992, p. 469) has suggested that the neocortex size and the number of neocortical neurons impose a cognitive upper limit on an individual's information-processing capacity. These limit "the number of relationships that an individual can monitor simultaneously" to around 150 nodes. Additionally, McCarty et al. (2001) use two methods to estimate the size of average networks. They use what they term the scale-up and summation methods, and the results show "a remarkable similarity between the average network size[s] generated by both methods (∼291)" (2001, p. 28). They estimate, however, that network sizes for various subpopulations can be substantially larger. These include clergy, politicians, labor organizers, and diplomats.

Sociolinguistic research has covered a part of the feasible network sizes. **Figure 1** visualizes this with the aid of dummy data. The x-axis indicates the size of networks and the y-axis the rate of innovation adoption for network types. The left-hand part shows the size of the networks covered, while the right shows how these fare with cognitively possible human network sizes.

We added a regression line to the visualizations but given the absence of empirical evidence it is impossible to know whether the line continues if we have evidence exclusively from small networks.

Recent findings from fields outside sociolinguistics suggest that network sizes play a more substantial role than previously thought. Ma et al. (2019) focus on trust in public and private social media groups, surveying 6,383 Facebook Groups users. Their observations show that people trust private groups more than they do public groups, which is to be expected. However, the differences between group types disappear once the group size exceeds circa 150 members. When networks become larger, individuals are no longer be able to perform the mental reasoning of who actually is in the group and who is not. Therefore, the difference between network types levels in large networks.

Moreover, increasing empirical evidence has recently led social network scholars to question the unidimensionality of the weak-tie model. Brashears and Quintane (2018) for instance elaborate on the idea of bandwidth in social contacts as an additional dimension. This concept refers to the total flow of information and accounts for capacity, frequency, and redundancy of network ties. Their model shows that even though humans acquire a smaller proportion of new ideas through strong contacts, the greater bandwidth of these contacts means that more total content is transmitted through these contacts. Strong contacts could therefore be more likely to transmit a greater share of novel information than weak ties, which could

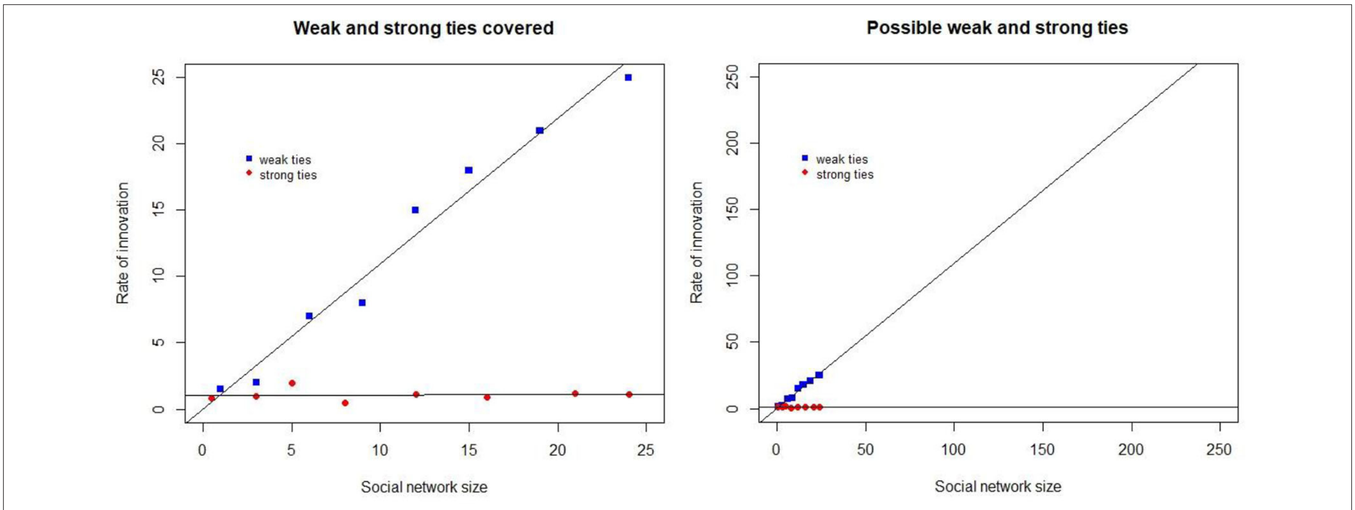


FIGURE 1 | A schematic representation of the network sizes covered (**Left**) and also the cognitively possible networks (**Right**).

explain the role of large urban working-class centers as places for innovation.

We investigate networks in Twitter and operationalize them using metadata available for each account. These are related to network size and mutual interaction patterns. Previous studies in computational sociolinguistics have used such information more to extract social network structures (Nguyen et al., 2016), but less to deepen understanding of the social network theory, which is the locus of this article. Eleta and Golbeck (2014) apply machine learning to study how social network characteristics and linguistic profiles influence language choice and how multilingual users of Twitter mediate between language groups in their social networks Their data consist of 92 ego networks, and the observations show that the proportion of English users in the network is the most significant predictor of language choice. Moreover, if a network consists of L2 users, this will increase the likelihood of L2 use. Kim et al. (2014) investigate how virtual networks impact multilingual practices, and they quantify "the degree to which users are the 'bridge-builders' between monolingual language groups." Hale (2014) studies networks utilizing mentions and retweets, and his results confirm the central role of multilingual users, and those who use English in particular, as the bridging forces in the network.

MATERIALS AND METHODS

To test the computational methods, we use two sets of Twitter data. Section A Cohort-Based Approach to Network Size uses evidence from the *Nordic Tweet Stream* corpus (NTS), which is a real-time monitor corpus of geolocated tweets and their metadata from the five Nordic nations (Laitinen et al., 2018). Section An Algorithmic Approach to Networks in Sociolinguistics utilizes an algorithm-based method, which makes use of mutual interaction data from a set of accounts from the Nordic region.

The NTS is being collected using the free Twitter Streaming API and the HBC (https://github.com/twitter/hbc) as the downloading mechanism. We apply a double filtering with the geolocation information and the Nordic country codes to ensure that the material originates from the region (Laitinen et al., 2018). While tweet data offer an efficient way of capturing big societal data, there are limitations. As an illustration, users who do not want to share their geolocation are not included. Depending on privacy settings and the geolocation method used, tweets either have (a) an exact location specified as a pair of latitude and longitude coordinates or (b) an approximate location specified as a rectangular bounding box. These geolocation data are available in the metadata attached to the message. Alternatively, no location at all is specified. For location, the data are derived either from the user's device itself (using the GPS) or by detecting the location of the user's Internet Protocol (IP) address (GeoIP). Exact coordinates are almost certainly from devices with built-in GPS receivers (e.g., phones and tablets). The GeoIP-based device location can be tricked by using proxy gateways. Attempting to hide one's location is probably most common amongst users with a malicious intent, such as bots.

To exclude bots and to increase data accuracy, we use a machine-learning algorithm developed by Lundberg et al. (2019). The version recognizes automatically generated tweets (AGTs) written in English and in Swedish. We define an AGT as a tweet in which all or parts of the natural language content are generated automatically by a bot or other type of program. The algorithm makes use of nine numerical and nominal properties that can be computed directly from the tweet metadata. The accuracy rate of the algorithm is over 97%. The results in section A Cohort-Based Approach to Network Size exclude possible bot accounts, whose share of AGTs is >50%, and section An Algorithmic Approach to Networks in Sociolinguistics focuses on genuine human accounts that have been selected manually.

The first method (based on cohorts) does not assume a pre-existing social network as the starting point but rather aims at

TABLE 1 | Raw statistics for the data used in section An Algorithmic Approach to Networks in Sociolinguistics.

Account	Friends	Net size	Loss rate (%)	Tweets	Retrieval (in mins)	Text collection (in mins)
account_01	409	221	46	312,350	230	38
account_02	335	166	51	253,758	181	33
account_03	309	195	37	286,945	201	33
account_04	332	175	47	150,774	184	25
account_05	201	105	48	100,915	105	14
account_06	418	132	68	192,944	140	23
account_07	468	281	40	316,944	291	41
account_08	448	286	36	322,566	303	40
account_09	418	216	48	189,628	229	26
account_10	496	297	40	516,686	282	67

uncovering quantitative patterns in the data. To measure network sizes and to correlate size with the rate of innovation, we use two metadata attributes available for each tweet. They measure the number of one's online friends and followers, and networks are operationalized as follows: The number of followers indexes truly weak ties (i.e., requires no action from a user), and the number of friends is an indication of slightly stronger links (i.e., requires user effort). We suggested previously that these metadata offer a way of measuring social networks and are ideal for research purposes, because they are automatically generated and hence they reduce the observer bias (Laitinen et al., 2017). They are also freely available to researchers with intermediate computing skills.

Similar to Milroy (1987), we operate under the assumption that social networks are abstractions, but we also propose that information from digital social network applications can be used to distinguish between ties of varying strengths. Friend and follower counts are useful indicators of social networks because of their differing qualities. Our definition of truly weak ties and slightly stronger ties is similar to Granovetter's (1973, p. 1361) assumption that the "strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." His methodology assumed stronger ties to be "friends," while weak ties consisted of "acquaintances," very similar to what we do below. By the same token, while we do not claim that friend count would indicate stronger ties in the sense in Milroy (1987), we assume that our operationalization of digital social networks is closely similar to the underlying idea of networks. Indeed, Milroy (1992, p. 178) argues that "a tie is 'weak' if it is less strong than the other ties against which it is measured," which also holds true for the follower counts when compared with friends.

The second method, the algorithm-based approach, zooms in on a set of real networks extracted by accessing account information through the Twitter API. We employ data such as friends and follower patterns, re-tweets, mentions, and directed messages. The accounts are anonymized, and we work with two types of network.

- Large (100–300 nodes) weak-tie networks
- Large (100–300 nodes) close-tie networks

We identified a set of accounts similar user profiles and extracted all interaction data available. The policy limitation of the API allows accessing up to 3,200 of the latest messages for each unprotected account. The account holders are from the metropolitan areas of Helsinki and Stockholm, are not working in academia, identify as males, have >10 messages primarily in English, and have more than 300 friends and followers. The last figure comes from a study that estimates median network sizes for multilingual individuals (Laitinen and Lundberg, 2020).

We narrowed the candidate accounts to ten and extracted their networks, including recent tweets and mutual interaction profiles. We excluded verified accounts (i.e., subpopulations with anomalous networks of politicians/celebrities/businesses) and accounts with more than 1,500 contacts (friends + followers). This was done to ease the time required for extracting mutual interaction data from large social networks. It is important to note that, while the number of accounts is small, the data extraction through the API takes circa 3–6 h per account (**Table 1**).

Even though the algorithm-based approach is tested with ten accounts, the size of our data is large. For instance, the mean network size is over 200 individuals (207), and the size of the textual data is over 2.6 million messages. In **Table 1**, the net size represents the number of collected accounts for the network (number of nodes in the graph). The loss ratio indicates the percentage of accounts lost after filtering.

The mutual interaction patterns are subjected to algorithms in order to assign labels of weak or strong networks to the accounts. The algorithms are explained in detail below, but they are mainly from the graph theory and the set theory, and some of them have been developed by us. For instance, we use betweenness centrality, which is a measure based on finding the shortest path between nodes (Freeman, 1977; Brandes, 2001) and closeness centrality (Perez and Germon, 2016). Kuikka (2018) argues that betweenness measures identify nodes that act as brokers between communities and are used to detect the density of how people are connected to each other in a network. We also use Jaccard Similarity Coefficient (JSC), which is a symmetric measure that calculates the similarity between two sets, and it is used to measure the similarity between accounts in terms of the number of common followers/friends. The assumption is that the share of common friends/followers is higher in a strong-tie network than in weak-tie settings. In addition, we assign weights to each account in the network and employ a method which we call disjointness. This last method enables us to estimate how well the nodes in a network are connected if the ego node were to be removed. The network labels are therefore multidimensional.

As for the dependent variables, we employ items that are frequent enough to be used in the testing phase. First, the cohort-based method uses the dominant language for each account. This information is available in the NTS metadata, and the share of English messages per account is correlated with network sizes. As our data come from the Nordic region, it ought to be noted that while English has no *de jure* position in the region, it is increasingly used as a lingua franca. Space does not permit us to discuss the sociolinguistic diversity of the region, but see country overviews in Modiano (2003), Preisler (2003), Leppänen et al. (2011), and Graedler (2014). Previous studies that use Twitter data have suggested that a great majority of messages in one location, a region for instance, are from residents of that location (Gonçalves et al., 2018; Lamanna et al., 2018) and not from visitors and tourists. The cohort-based method uses information from tens of thousands of accounts, and we assume that our dataset is reliable, given the general limitations of Twitter data. We use automatically-assigned language labels, and although automated language identification methods are not entire accurate, the agreement between human coders and Twitter's language recognition system is fairly high for languages written in the Latin alphabet (Graham et al., 2013).

Second, the algorithmic approach uses a mixture of linguistic features available in the tweet text. These features consist of contracted forms (*won't*, *'ll*, *I'm* etc.), and *NEED* to used as a semi-modal auxiliary. These features are qualitatively different as the contracted forms index colloquial, spoken-like use (Biber et al., 1999, p. 1128–1132), while *NEED to* is currently undergoing change in English (Leech, 2013) and is highly pervasive in ELF use in the Nordic region (Laitinen, 2016).

RESULTS

A Cohort-Based Approach to Network Size

We illustrate the cohort-based method first using data from 199,832 accounts from the NTS, from which we removed subpopulations with anomalous network profiles, as defined in section Social Networks in Variationist Sociolinguistics. After the initial results, we test the findings with data from which software bots are removed. These bot-free data consist of 90,887 accounts, obtained from the NTS but limited to Sweden only (labeled as NTS-Human-Swe).

The null hypothesis is that increasing the number of network ties does not lead to increases in the share of English per account. The cohort-based approach for both categories is specified in (1)–(6) (it refers to followers in the NTS, but the same procedure applies to friends and to both datasets):

(1) We sort out all the accounts based on their followers' counts.

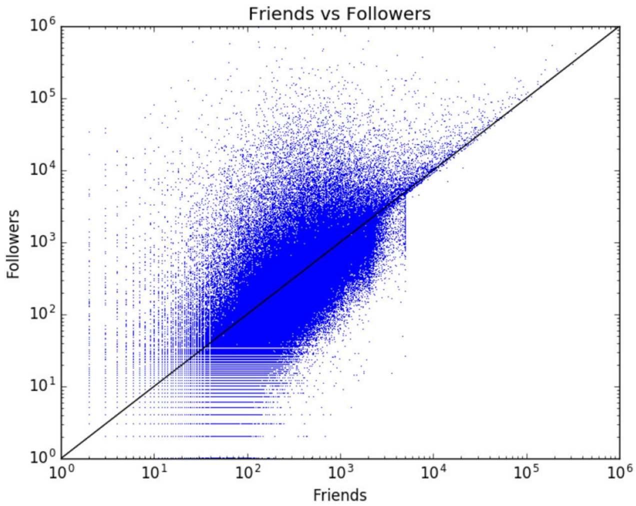


FIGURE 2 | Friends and followers visualized (199,832 accounts).

(2) The accounts are divided into N equally-sized cohorts where cohort 1 is the 199,832/N, and it has the lowest follower count, and cohort N has the highest. N can of course be adjusted.
(3) We compute the percentage of tweets written in English per each account.
(4) The language identifier used is Twitter's own language identification tool, the accuracy of which is discussed in section Material and Methods.
(5) We can adjust the proportions of English in the tweet stream (EngMajor) for each cohort and associate the cohorts with the EngMajor percentage. The results here use >50% share of messages in English (for other proportions, see Laitinen et al., 2017).
(6) We correlate the cohorts against the percentages and visualize them.

An average account profile in the NTS is such that the median size of networks is 235 friends and 195 followers. **Figure 2** shows how the friend and follower counts are distributed in the data. There is a relatively straightforward (x = y) spread of the values. The only exception is the friends category, in which Twitter imposes an upper limit of 5,000 friends that each individual account can follow (https://support.twitter.com/articles/66885#). The only way to increase one's friends count is to gain new followers, and therefore there is an even more direct correlation of friends/followers after the 5,000 mark.

Figure 3 (left) illustrates a 10-cohort division visualizing how cohorts differ in terms of the >50% percent threshold. The result shows that more Twitter followers means more messages in English, with the non-parametric Kendall tau correlation coefficient (0.956) indicating a strong positive correlation between the two vectors at statistically significant levels ($p < 0.0001$). Note that cohorts 1–4 are accounts with fewer than the median number of followers (i.e., 195).

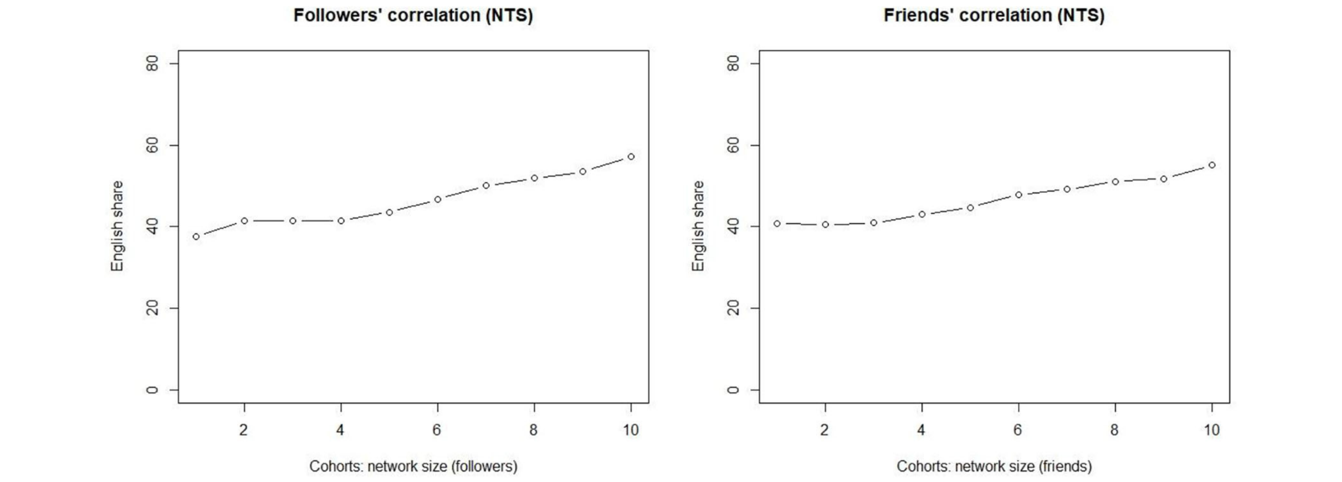


FIGURE 3 | The correlation between followers (**Left**) and friends (**Right**) and the share of accounts in which English dominates.

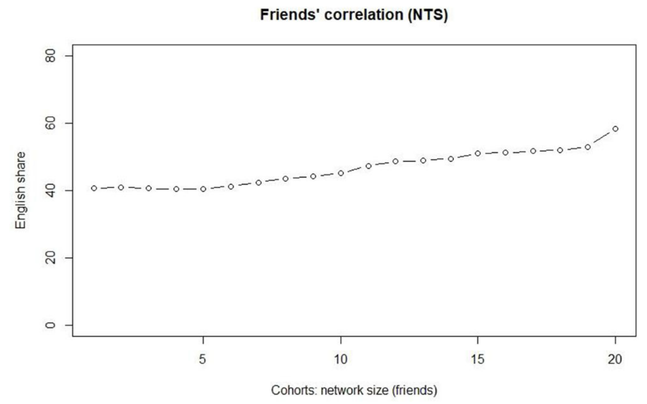


FIGURE 4 | The correlation between 20-cohort friend category and the EngMajor.

The quantitative pattern with these truly weak ties is clear. The correlation between the follower counts and the use of English is linear, and the correlation is strong. Of the accounts in which the number of followers is lower than the median, roughly 40% have the majority of their messages in English. The higher that we move in the cohorts, the higher is the share of such accounts. At the other extreme, in the cohorts with the highest number of followers over half of the accounts fulfill the criterion.

The quantitative pattern for the slightly stronger ties (friends) is shown on the right. The correlation between the number of friends and the increase in the use of English is strongly positive, with the Kendall tau correlation coefficient at 0.867 ($p < 0.0001$), i.e., for all of the 199,832 accounts in the dataset, more online friends means a larger share of messages in English.

However, contrary to what is observed with truly weak ties, the stronger network index behaves differently. For small networks, the increase in network size has no impact on the response variable. It is only from cohort 4 onwards that the share of EngMajor increases when we increase the number of friends in the network.

These results suggest that there is a straightforward correlation in the truly weak tie networks, but the friend data indicates that the distinction between weak ties and stronger ties levels out when the network size is large enough. If we had restricted our analysis only to traditional small networks of 30–50 nodes in ethnographic attempts, our data would have confirmed the customary finding related to the diffusion of innovations and network strength. That is, weak ties promote change, and stronger ties prevent it. However, the results obtained using this approach suggest that this is not necessarily the case. Once the network size grows to become large, the traditional distinction between weak and stronger ties disappears. Note that we are not referring to the percentages of the accounts, but to correlational patterns of the variable. Large networks here mean that the network sizes are still within the cognitive limits (see section Material and Methods).

To explain this finding, we must balance between the limitations and the advantages of our data. The most obvious limitation is that we might observe a random quantitative pattern that emerges from messy data. Moreover, we do not know anything about the density or the multiplexity of the network ties but can only assume that the friends category represents a slightly stronger network index, since it involves an active decision to follow someone. The friends network index might also include a greater share of interactive networks. To tackle the limitations, the next section applies a different method and approaches ego networks.

The obvious advantage is the size of our data. Each cohort in **Figure 3** consists of nearly 20,000 accounts, and we are not restricted to small ethnographic records. The network size for the first three cohorts is 0–122. As pointed out earlier, the median number of friends is 235. The results support rejecting the null hypothesis, but the threshold level of 122 stems from an arbitrary value of ten cohorts.

Figure 4 tests the observations using 20 cohorts. As the interest is on slightly stronger ties, we only use the friends data. The figure confirms the observation and indicates a leveled proportion of EngMajor for the first five cohorts. After that, the network size correlates positively with the increasing use of English in the tweet stream. The Kendall tau correlation coefficient is 0.905 at a statistically highly significant level ($p < 0.0001$).

Cohorts 1–5 consist of networks of <100 individuals, and a marked increase takes place only after cohort 5 (100–122 individuals). The share of accounts with a >50%+ share of tweets in English increases systematically for each cohort so that for cohort 6 it is 41.2%, and for cohort 19 it is 51.9%. Cohort 20 has its friends count at over 1700, and according to our present understanding, these represent "evangelists" in the Krishnamurthy et al. (2008) sense, i.e., they are more or less automated bots aiming at increasing their friends basis automatically.

Figure 4 suggests that the threshold network size after which the distinction between weak ties and slightly stronger ties levels is of around 122 nodes. Next, we zoom into the bot-free data, and the main question is whether we can replicate the findings using the bot-free data. Overall, the number of bots in the Swedish subset is low (1,149 accounts = 1.0%), but they generate a high number of tweets (404,804 = 7.6%). The majority language in the bots is English, since nearly 20% of all of the English tweets were identified as AGTs, but the corresponding share for Swedish was <2% (see Laitinen and Lundberg, 2020). The visualizations also exclude the smallest networks of fewer than five nodes.

The bot-free quantitative patterns are shown in **Figure 5**, and they are similar to those observed earlier. As for followers (left), they show a linear increase in the share of messages in the English per cohort as we move to the right on the x-axis. The correlation between network size and the share of English is not only straightforward but also statistically significant, as the Kendal tau correlation coefficient is one ($p < 0.0001$). For smaller networks, the share of English is around 40%, and it increases for every increase in the network size, so that the share for the largest networks is well over 50%. The increases are slight, but the shares of the English use nevertheless increase for each cohort. Once the network size grows larger, we observe more noteworthy increases.

The right-hand side visualizes the slightly stronger ties (friends) and verifies the initial observations. These results confirm the findings presented above. The observations show that the correlation with slightly stronger ties is equally linear, and this is also supported by the Kendal tau value (0.944, $p < 0.0001$). However, the share of English actually decreases for the small stronger-tie networks. That is, the empirical evidence presented here suggests that truly weak ties and slightly stronger ties behave slightly differently for small networks, but the distinction disappears once the network size grows larger. The share of English remains flat for cohorts 1–3 of the truly weak ties (left), while the share actually decreases for the slightly stronger ties for the smallest networks (right). Cohort 4 consists of those whose network size exceeds 120 nodes.

The present section has presented our cohort-based approach to measuring networks in social media. While we acknowledge that the method is straightforward, it has obvious benefits for this type of big and rich data approaches to language variability and social networks. The method is light in terms of computing power, as the values can be easily obtained from the data stream. In addition, we can use data in their entirety since each account makes the values directly available with minimal or no data loss.

The obvious difference between this approach and the ethnographically-oriented data-collection in Milroy (1987) is that our method does not deal with ego networks but rather takes a top-down approach, correlating network size and a linguistic feature. As for the innovation, previous studies have shown that English in the Nordic region is closely associated with age; this means that the younger generations clearly use English as an additional tool more often than do the older groups (Leppänen

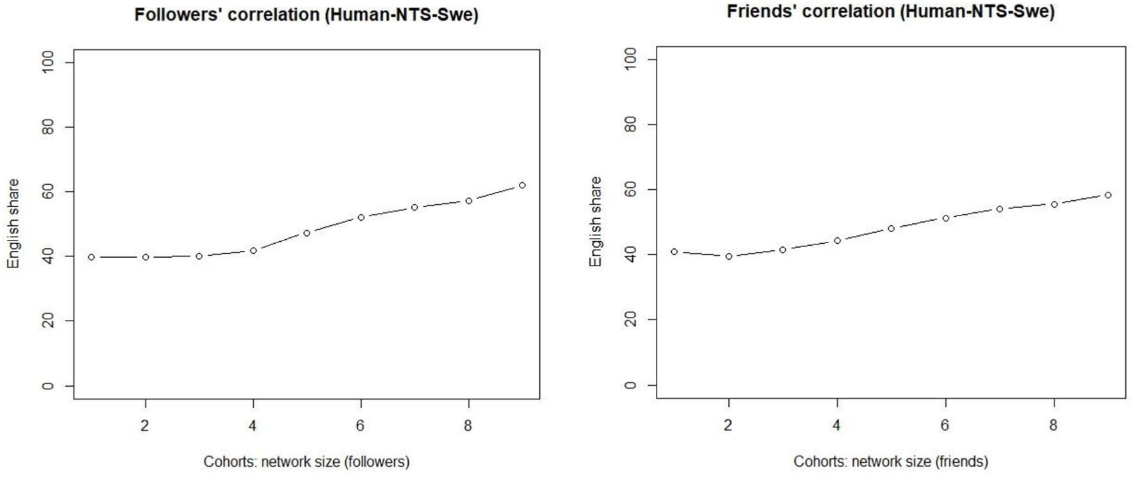


FIGURE 5 | Bot-free correlations of truly weak ties (**Left**) and slightly stronger ties (**Right**).

et al., 2011). Unfortunately, age is not included in the metadata parameters in the raw data, and its role cannot be controlled.

The main finding here is that we can confirm our pilot results in Laitinen et al. (2017). The new cohort-based findings using bot-free data suggest that network size plays a role in leveling the differences beyond a certain threshold. The following section will turn its attention to ego networks.

An Algorithmic Approach to Networks in Sociolinguistics

This section digs deeper into digital networks and uses an algorithmic method that complements the results above and provides tools for analyzing networks of mobile individuals. We operate with the 1-step neighborhood, which consists of a focal node, ego, and nodes directly connected to it. We also include the connections between nodes (degree 1.5). Twitter is a directed-graph network, and we are interested in what accounts "see" instead of how they are "seen," and consider friends rather than followers in the analysis. Consequently, we deal with a graph-based structure in which nodes represent accounts and directed edges are considered as a friend relationship, as in **Figure 6**, which visualizes two nodes in which A is either following B, or B is a friend of A.

The method assumes that account activities and mutual interaction between accounts have an impact on the relationship. To subject activities to the algorithms, we collected up to 3,200 recent tweets in JSON files for each account in the network and then extracted the values for how many times accounts in the entire network retweet or quote another account in the same network, and counted the number of times that accounts mention each other.

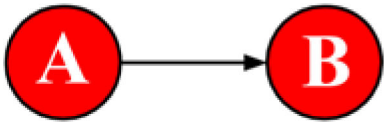


FIGURE 6 | A simplified example of a directed graph.

In order to extract ego-networks and to assess network values (either weak-tie and close-knit), we applied multiple criteria to the edges and nodes. While many of them are used in data mining, they measure network activities rather like the ethnographic methods in Milroy (1987) but applied to the parameters available in digital social networks.

First, we use a linear combination in (1), in which we assign weights to the links in the network.

$$\begin{aligned} \textit{Edge weight} \;=\; & (w_1 * \textit{retweet}_{\textit{count}}) + (w_2 * \textit{quote}_{\textit{count}}) \\ & + (w_1 * \textit{mention}_{\textit{count}}) \end{aligned} \qquad (1)$$

Where w_1, w_2, and w_3 are weights that can be assigned based on the application of interest so that $\sum_{i=1}^{3} w_i = 1$. Weights regulate the importance of each feature in the analysis. For instance, if we want to focus on the number of retweets, we assign $w\_1 = 1$ while w_2, $w\_3 = 0$. Moreover, we assume that those accounts that have a higher rate of publishing tweets have more impact on the information flow in a network, which should be considered as a factor. The point is to separate active accounts from those that use Twitter passively while rarely creating any content. To assign weights, we extracted the age (in days) of each account and the total number of tweets. Then, calculating the average number of tweets per day for each account and using (2), we can assign weights to the individual nodes as well.

$$\textit{Node weight}\ (A) = \frac{\textit{average tweets per day for account A}}{W}, \qquad (2)$$

$$\textit{where}: \quad W = \sum_{i=1}^{N} \textit{average tweets per day for account } A_i. \qquad (3)$$

Figure 7 visualizes an ego network with 30 nodes and 142 edges, (a) without assigning weights to the nodes and edges, and (b) by assigning weights using the formulae in (1)–(3). The larger the node, the higher the value for tweets per day, and the thicker the link, the stronger the connection between the nodes.

Second, we use *betweenness centrality* (BC) to detect the density and to interpret how people in a network are connected

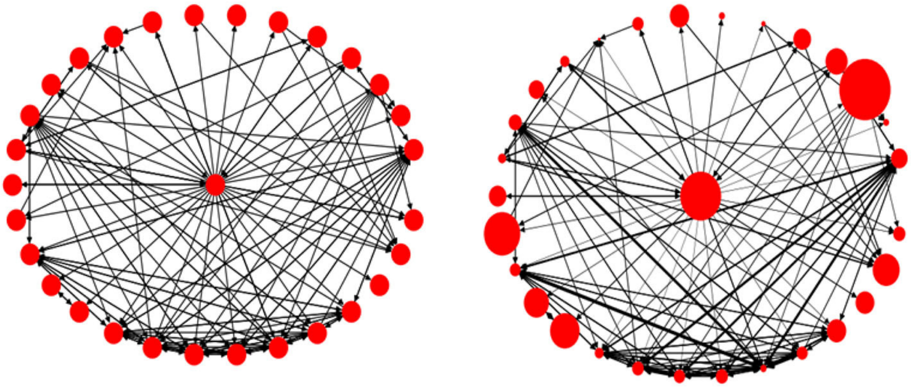


FIGURE 7 | An ego network, without assigning weights (**Left**), and with weights (**Right**).

to each other. The BC values represent the ratio with which an account establishes the shortest path between any pair in the network (Freeman, 1977). In other words, the BC of node v is the sum of the fraction of all of the shortest paths for any pair of nodes in the network that pass through v:

$$C_B(v) = \sum\nolimits_{s,t\,\in\,V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{4}$$

Where V is the set of nodes in the network, $\sigma(s,t)$ is the number of shortest paths between nodes s and t, and $\sigma(s,t|v)$ is the number of shortest paths between s and t that pass through V. Hence, the lower the BC value, the fewer the shortest paths passing through that account, and *vice versa*. The assumption is that the lower the spread (i.e., the difference between the higher and the lower values) of BC values in a network, the more connected the accounts are to each other, and the network consists of strong ties.

Consider **Figure 8**, in which the spread of the BC values is zero. The network is complete as all the nodes are connected to each other and the shortest path between each pair of the nodes is the direct path between those two nodes, and the path does not pass through any other nodes.

Figure 9 visualizes two real Twitter networks. The yellow nodes represent the ego, while the black links represent Two-way connections and blue links show One-way connections. Using visual cues, we can see that the left side is a weak-tie network, while the one on the right represents a stronger-tie network, and this is also supported by quantitative evidence. The spread value for the weak-tie network is 0.5455 and the corresponding value for the strong ties is 0.3014. We use normalized BC values to address the effect of network sizes on the calculations.

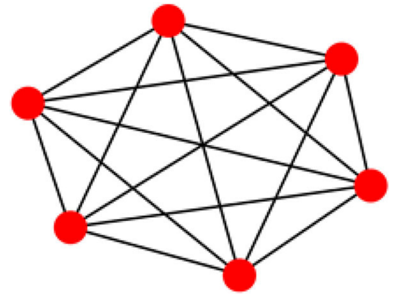


FIGURE 8 | A complete graph with 6 nodes and BC mean and spread.

The third measure is *closeness centrality* (CC), a concept that measures the distance between nodes (Perez and Germon, 2016). In the graph theory, the distance between two nodes is defined as the length of the shortest path between two nodes. CC is the reciprocal of the sum of the distances from a node to all the other nodes in the network. As in the case of the BC analysis, to eliminate the effect of network size we applied the normalized CC values in the analysis. The normalized CC value is calculated using the formula in (5):

$$C_C(v) = \frac{N-1}{\sum_{i=1}^{N-1} d(u,v)}. \tag{5}$$

Here, $d(u,v)$ is the shortest-path distance between u and v, and N is the number of nodes in the network. The CC values are between 0 and 1 for each node, and higher values of closeness on average could be interpreted as higher connection rates between nodes. In a directed graph in Twitter, there are two CC values for each node (i.e., incoming and outward). If the difference between the two CC values on average is low, it indicates that the majority of the connections in a network are Two-way links. Therefore, the network is a stronger-tie network.

The next two measures have been purpose-built by us and can be illustrated by inspecting the two networks in **Figure 9**, above. In the weak-tie network (left), the majority of the accounts are connected to each other through the ego node, while the accounts in the right-hand network are not only connected to the ego node but to the other accounts in the network as well, which means that the network consists of stronger ties. If we remove the ego node and its incoming/outgoing links from the data, we can then calculate the ratio of *disjoint nodes* in the network. We assume that the higher the value of the disjointness ratio, the weaker the network will be. Furthermore, as mentioned before concerning the edge weights, we can calculate the mean values of the edge

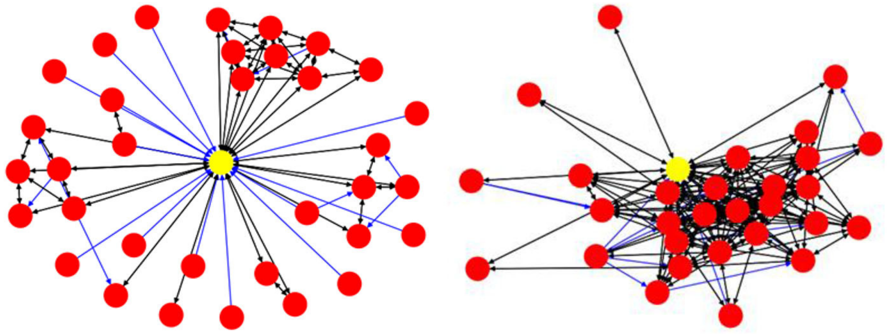


FIGURE 9 | A weak-tie ego network (**Left**) and a strong-tie network (**Right**).

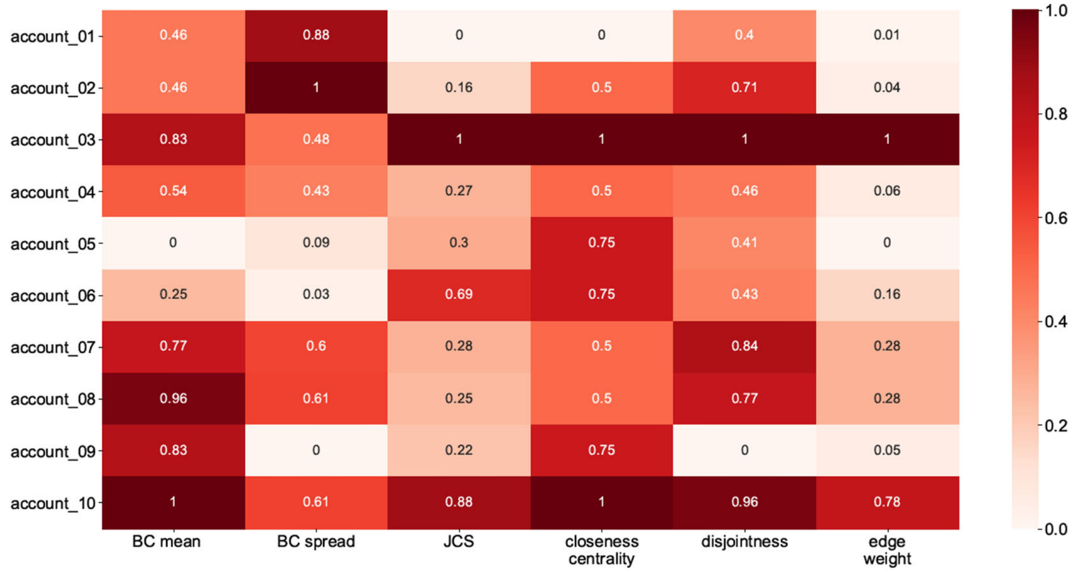


FIGURE 10 | Ten candidate accounts and their corresponding values for indices.

weights for each network. We would argue that, for a stronger-tie network, the mean value of the edge weights should be higher than the corresponding value for a weaker-tie network because individuals in a strong-tie network might be expected to have more interaction and activities than in a weaker-tie network.

The last algorithm strengthens the method by bringing in a tool that enables us to measure the similarity between two sets. It builds on the idea that individuals in a strong-tie network might be expected to be more similar to each other than individuals in a network characterized by weak ties. If we use Milroy's (1987) ethnographic work as our point of comparison, men in the Belfast neighborhoods were localized and spent more time with those who were similar to themselves in their dense strong-tie networks than women.

To measure similarity between sets, we use the Jaccard Similarity Coefficient (JSC). It is a symmetric measure that can be used to calculate the similarity between sets A and B as follows:

$$JSC = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

The assumption is that if two accounts have a high number of shared friends (i.e., a high JCS value), they are more similar to each other than two other accounts with a lower JCS value. Consequently, if the average JCS values for all the nodes in ego network A are higher than the averages for another network B, it means that the accounts in the A network are more similar to each other and that we are dealing with a stronger-tie network, and *vice versa*.

Consider the two networks presented in **Figure 9**, above. Using the formula presented in (6), we can calculate the mean JSC value for the weak-tie network to be 0.12 and the corresponding value for the stronger-tie to be 0.9. The average similarity for the network on the right is almost 8 times higher than the average similarity for the network on the left.

To measure the network qualities, we extracted the values for each network and, with the aid of Min-Max normalization, placed them on an interval [0,1]. We subtracted the calculated values for the BC mean, BC spread, disjointness ratio, and CC difference from 1 in order to make them comparable with the other features. The values are shown in **Figure 10**. The higher values for each feature (i.e., the darker the cell) indicate stronger-tie networks, and *vice versa*.

To assign labels (weak-tie or strong-tie) to the candidate networks, we calculated the mean values (strength coefficient *alpha*) for each cell in **Figure 10**. We then labeled the accounts with lower alpha values as weak-tie networks (W1–5) and the rest as strong-tie networks (S6–10), as shown in **Figure 11**.

The strength values (top) and the visualizations of all of the ten networks suggest that the algorithms are able to distinguish between networks with differing qualities. The visualization shows that the candidate networks as a whole can be roughly divided into weak-tie networks and networks with stronger ties. The method is robust and is not affected by smaller clusters that might appear, for instance, inside a weak-tie network. As a whole, therefore, we are able to suggest that the differences between the network types are supported by complex multidimensional quantitative data and visual cues. The next step is, then, to test to see whether differing network structures are reflected in the linguistic behavior.

In the last part of this study we investigate how the dependent variables, listed in section Materials and Methods above, are distributed among the network types. The accounts, their sizes, and the normalized frequencies (per 100,000 messages) of the

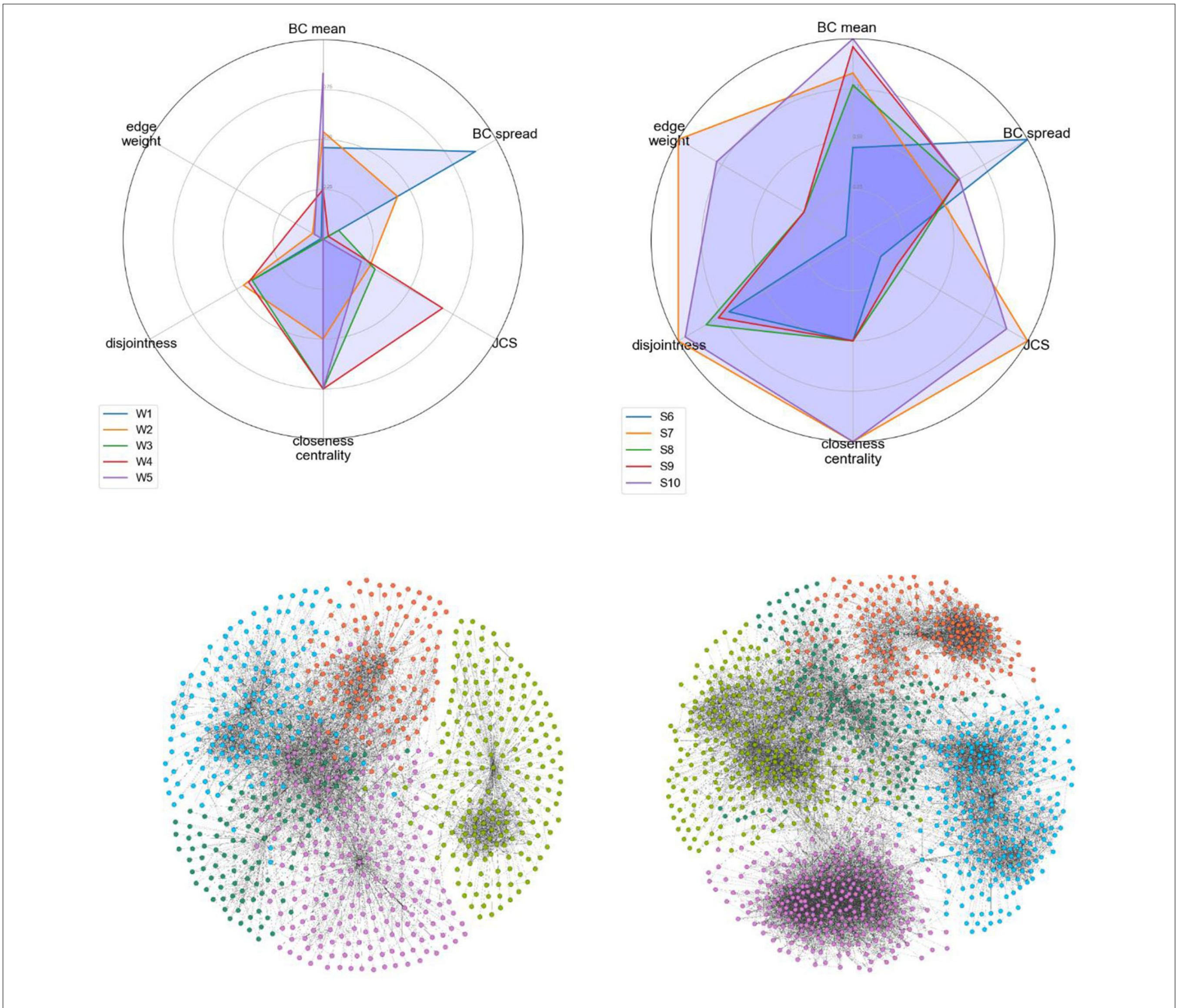


FIGURE 11 | Visualizing all the candidate networks.

dependent variables are shown in **Table 2**, below. The three columns on the right show the number of English messages in the network, the number of contractions in the text, and the frequency of *NEED to* + V constructions. It is important to note that, while the observations are based on a limited number of accounts, the data have been retrieved from the entire network connected to the ego node. These data consist of a total of 2,074 network nodes with over 2.6 million messages and nearly 30 million tokens of text. The network sizes vary, with the smallest possessing 105 nodes and the largest nearly 300. The number of messages varies between 100,915 and over half a million. The mean is 264,351 messages.

Figure 12 shows three boxplots that visualize the relationships between the weak- and strong-tie networks and the three

TABLE 2 | Statistics related to the dependent variables (normalized per 100,000).

Account	Network	N msg.	EngShare	Contr.	*NEED to* + V
W1	221	312,350	63,220	3,860	630
W2	175	150,774	40,910	940	310
W3	105	100,915	81,195	3,580	770
W4	132	192,944	45,237	3,230	560
W5	216	189,628	68,688	3,800	610
S6	166	253,758	79,534	2,590	840
S7	195	286,945	61,039	2,930	660
S8	281	316,944	85,387	5,790	890
S9	286	322,566	62,170	2,610	450
S10	297	516,686	81,261	6,290	1,070

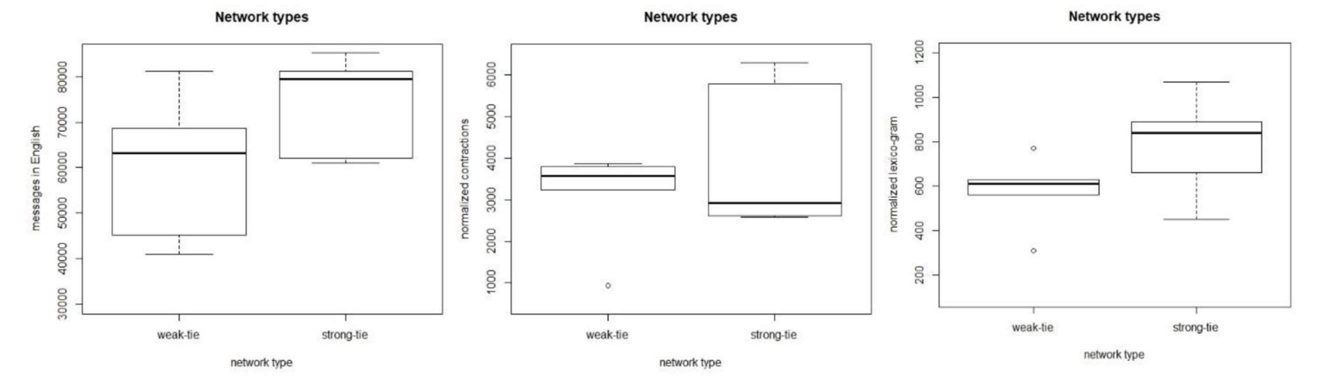


FIGURE 12 | The relationships between the network types and the dependent variables.

dependent variables. The data show no consistent pattern in which large networks would be quantitatively different from each other, but large weak and strong-tie networks behave similarly in terms of these variables. For the count of English messages (left), the mean value for the strong-tie networks is higher, but when tested with the Welch Two Sample *t*-test for independent samples, the differences between the networks are not statistically significant ($t = -1.55$, $p > 0.05$). The mean value for the contracted forms is slightly higher for the weak-tie networks, but the differences are not statistically significant ($t = -0.97$, $p > 0.05$). As for the lexico-grammatical variable, the mean is higher for the strong-tie networks, but the differences are not statistically significant ($t = -1.55$, $p > 0.05$).

The quantitative patterns observed are clear. When we investigate the large networks whose sizes are above the threshold level suggested in section A Cohort-Based Approach to Network Size, we can observe identical patterns. The results show no distinction between large weak-tie and strong-tie networks, which suggests that the differences observed in small ethnographic studies level out when the network size becomes sufficiently large. These observations support the cohort-based findings in section A Cohort-Based Approach to Network Size, above, and they also introduce ways of measuring the digital networks of mobile individuals in the social media.

We have attempted to demonstrate our algorithmic method which utilizes data-mining of the social media and uses a range of quantitative measures to establish network indices. The method enables us to establish networks of varying strengths and to determine that these varying qualities can not only be visually confirmed (**Figure 11**) but also supported by quantitative information. The method requires some computational power but still involves a qualitative element, since we have endeavored to ensure that the candidate networks represent similar content profiles. As we point out above, previous studies have suggested that various subpopulations have anomalously high network profiles (McCarty et al., 2001), and, at this stage, the objective has been to ensure that the candidate networks are similar. Our future objective is to test the algorithmic method with a far larger set of networks.

CONCLUSIONS

This article has investigated digital social networks of highly mobile individuals, and we have attempted to contribute to the study of social networks in sociolinguistics by providing tools for accessing large networks. The research objective has focused on the role played by network size as a key determinant in social networks. We have shown that network size has not been used in variationist sociolinguistics. Recent network studies in other fields have, however, suggested that network size could play an important role and that the distinction between network types might level out beyond a given threshold size of networks (Ma et al., 2019). Another of our motivations has been to observe real networks whose size is close to the average (at least in Western societies). The mean size of the ego networks (207 nodes) used in section An Algorithmic Approach to Networks in Sociolinguistics far exceeds the size of networks that have been covered in previous sociolinguistic studies, but they still fall within the limits of viable networks, as discussed in section Social Networks in Variationist Sociolinguistics.

As for the research questions, the first question focused on improving the methods used in sociolinguistics so that the quantitative variable of a network could be better operationalized in situations where the population consists of both socially and geographically highly mobile individuals. We have introduced two methods for accessing the networks of mobile individuals, thus expanding the empirical basis from small-scale ethnographic observations. Section A Cohort-Based Approach to Network Size introduced cohort-based methods, while in section An Algorithmic Approach to Networks in Sociolinguistics we detailed an algorithmic approach. The methods have a strong empirical basis, and they offer new tools for variationist sociolinguistics. They reveal fundamental differences in comparison with ethnographic approaches. For instance, one of the advantages of ethnographic social network studies is that the methods build on the idea that networks are intrinsically a participant-related concept rather than something than an outsider analyst could construct (Milroy and Llamas, 2013). Our cohort-based method adopts an alternative approach, a clearly analyst-driven approach aimed at uncovering broad quantitative patterns in data rather than looking at existing networks. However, the algorithmic approach is very similar to the original idea, since the starting point is an existing network. As in Milroy and Milroy (1978) and Milroy (1987), the second method assumes the unit of study to be essentially a pre-existing category. Moreover, our method assumes network ties to be multidimensional, as the algorithms account for not only frequency of communication, but also a range of other factors. This means modernizing the network concept in sociolinguistics and bringing it closer to the contemporary idea that networks are not based on a simple dichotomy but consist of a range of attributes (Brashears and Quintane, 2018).

The second research question concentrates on the effect of network size on the validity of the theory by combining methods from sociolinguistics with computer science. Our results gained from both methods suggest that network size plays a role, and that the distinction between weak ties and stronger ties levels out once the network size grows beyond roughly 120 nodes. This finding is similar to the finding related to trust in networks (see section Social Networks in Variationist Sociolinguistics, above). We would, therefore, suggest that further studies be made of the digital networks of mobile individuals. Our raw data and the code are publicly available to other researchers.

Our future plans include continuing to work using the two methods. We plan to expand the cohort-based method and to test it with other dependent variables than simply language choice. Moreover, the metadata available in the tweet stream contain a number of possible predictors other than network size, and they need to be tested using linear regression. As for the algorithmic approach, our objective is to collect data from (tens of) thousands of accounts to scale up the method.

DATA AVAILABILITY STATEMENT

The Twitter dataset used in section A Cohort-Based Approach to Network Size is publicly available through the streaming API (https://developer.twitter.com). The data used in section An Algorithmic Approach to Networks in Sociolinguistics can be made available by the authors, without any undue restrictions, to qualified researchers. The code used in the algorithmic approach is available through GitHub (https://github.com/Masoud-Fatemi/Two-approaches-to-digital-social-networks).

AUTHOR CONTRIBUTIONS

This study was conceptualized by ML and MF. JL was responsible for data curation together with MF. The investigations were carried out by ML and MF. The methodology developed by MF, JL, and ML. The visualizations were created by MF and ML. The project was administered by ML, who was also responsible for writing the original draft version. All authors contributed to the article and approved the submitted version.

FUNDING

This research has been supported by the generous funding of the Center for Data Intensive Sciences and Applications (DISA) at Linnaeus University. In addition, ML has been able to invest research time offered by his position at the University of Eastern Finland.

ACKNOWLEDGMENTS

We wish to thank Professor Emerita Lesley Milroy for her valuable comments on this article. The usual disclaimers apply.

REFERENCES

Aral, S., and Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *Am. J. Sociol.* 117, 90–171. doi: 10.1086/661238

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Mathem. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Brashears, M. E., and Quintane, E. (2018). The weakness of tie strength. *Soc. Networks* 55, 104–115. doi: 10.1016/j.socnet.2018.05.010

Chambers, J. K. (2003). *Sociolinguistic Theory. Linguistic Variation and its Social Significance. 2nd Edn*. Oxford: Blackwell.

Coats, S. (2017). "European language ecology and bilingualism with English on Twitter," in *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities*, eds C. Wigham and E. Stemle (Bozen/Bolzano: Eurac Research), 35–38.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* 22:6, 469–493. doi: 10.1016/0047-2484(92)90081-J

Eleta, I., and Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Comp. Hum. Behav*. 41, 424–432. doi: 10.1016/j.chb.2014.05.005

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40, 35–41. doi: 10.2307/3033543

Gonçalves, B., Loureiro-Porto, L., Ramasco, J., and Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PLoS ONE* 13:e0197741. doi: 10.1371/journal.pone.0197741

Graedler, A-L. (2014). Attitudes towards English in Norway: a corpus-based study of attitudinal expressions in newspaper discourse. *Multilingua* 33, 291–312. doi: 10.1515/multi-2014-0014

Graham, M., Hale, S., and Gaffney, D. (2013). Where in the world are you? Geolocation and language identification in Twitter. *Profes. Geogr.* 66, 568–578. doi: 10.1080/00330124.2014.907699

Granovetter, M. (1973). The strength of weak ties. *Am. J. Sociol.* 78, 1360–1380. doi: 10.1086/225469

Granovetter, M. (1983). The strength of weak ties: a network theory revisited. *Sociol. Theory* 1, 201–233. doi: 10.2307/202051

Hale, S. (2014). "Global connectivity and multilinguals in the Twitter network," in *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto), 833–842.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Comp. Environ. Urban Syst.* 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Kim, S., Weber, I., Wei, L., and Oh, A. (2014). "Sociolinguistic analysis of Twitter in multilingual societies," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago), 243–248.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). "A few chirps about Twitter," in *Proceedings of the first workshop on Online social networks (WOSN '08)*, eds. C. Faloutsos, T. Karagiannis and P. Rodriguez (New York, NY: ACM), 19–24. doi: 10.1145/1397735.1397741

Kuikka, V. (2018). Influence spreading model used to analyse social networks and detect sub-communities. *Comp. Soc. Netw.* 5:12. doi: 10.1186/s40649-018-0060-z

Labov, W. (2001). *Principles of Linguistic Change. Vol. 2. Social Factors*. Oxford: Blackwell.

Laitinen, M. (2016). "Ongoing changes in English modals: on the developments in ELF," in *New Approaches in English Linguistics: Building Bridges,* eds O. Timofeeva, S. Chevalier, A.-C. Gardner, and A. Honkapohja (Amsterdam: John Benjamins), 175–196.

Laitinen, M., and Lundberg, J. (2020). "ELF, language change and social networks: evidence from real-time social media data," in *Language Change: The Impact of English as a Lingua Franca,* eds A. Mauranen and S. Vetchinnikova (Cambridge: Cambridge University Press).

Laitinen, M., Lundberg, J., Levin, M., and Lakaw, A. (2017). "Revisiting weak ties: using present-day social media data in variationist studies," in *Exploring Future Paths for Historical Sociolinguistics,* eds T. Säily, M. Palander-Collin, A. Nurmi, and A. Auer (Amsterdam: John Benjamins), 303–325.

Laitinen, M., Lundberg, J., Levin, M., and Martins, R. (2018). "The Nordic Tweet Stream: a dynamic real-time monitor corpus of big and rich language data," in *Proceedings of Digital Humanities in the Nordic Countries 3rd Conference* (Helsinki). Available online at: CEUR-WS.org/Vol-2084/short10.pdf (accessed April 11, 2019).

Lamanna, F., Lenormand, M., Henar Salas-Olmedo, M., Romanillos, G., Gonçalves, B., and Ramasco, J. (2018). Immigrant community integration in world cities. *PLoS ONE* 13:e0191612. doi: 10.1371/journal.pone.0191612

Leech, G. (2013). "Where have all the modals gone? An essay on the declining frequency of core modal auxiliaries in recent standard English," in *English Modality: Core, Periphery and Evidentiality,* eds J. I. Marín-Arrese, M. Carretero, J. Arús Hita, and J. van der Auwera (Berlin: Mouton de Gruyter), 95–115.

Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., et al. (2011). *National Survey on the English Language in Finland: Uses, Meanings and Attitudes. (Studies in Variation, Contacts and Change in English, 5)*. Available online at: http://www.helsinki.fi/varieng/journal/volumes/05/ (accessed April 11, 2019).

Lippi-Green, R. (1989). Social network integration and language change in progress in a rural alpine village. *Lang.Soc.* 18, 213–234. doi: 10.1017/S0047404500013476

Lundberg, J., Nordqvist, J., and Laitinen, M. (2019). "Towards a language independent Twitter bot detector," in *Proceedings of the Digital Humanities in the Nordic Region (DHN2019)* (University of Copenhagen). Available online at: http://ceur-ws.org/Vol-2364/28_paper.pdf (accessed April 11, 2019).

Ma, X., Cheng, J., Iyer, S., and Naaman, M. (2019). "When do people trust their social groups?," in *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (New York, NY: ACM).

McCarty, C, Killworth, P., Bernard, R., Johnsen, E. C., and Shelley, G. A. (2001). Comparing two methods for estimating network size. *Hum. Organ* 60, 28–39. doi: 10.17730/humo.60.1.efx5t9gjtgmga73y

Milroy, J. (1992). *Linguistic Variation and Change*. Oxford: Blackwell.

Milroy, J., and Milroy, L. (1978). "Belfast: change and variation in an urban vernacular," in *Sociolinguistic Patterns in British English,* ed P. Trudgill (London: Edward Arnold), 19–36.

Milroy, J., and Milroy, L. (1985). Linguistic change, social network and speaker innovation. *J. Linguist* 21, 339–384. doi: 10.1017/S0022226700010306

Milroy, L. (1987). *Language Change and Social Networks. 2nd Edn.* Oxford: Blackwell.

Milroy, L., and Llamas, C. (2013). "Social networks," in *The Handbook of Language Variation and Change,* eds J. K. Chambers and N. Schilling (Oxford: Blackwell), 407–427.

Milroy, L., and Milroy, J. (1992). Social network and social class: toward an integrated sociolinguistic model. *Lang. Soc.* 21, 1–26. doi: 10.1017/S0047404500015013

Modiano, M. (2003). Euro-English: a Swedish perspective. *Eng. Today* 19, 35–41. doi: 10.1017/S0266078403002074

Nevalainen, T. (2000). Mobility, social networks and language change in early modern England. *Eur. J. Eng. Stud.* 4, 253–264. doi: 10.1076/1382-5577(200012)4:3;1-S;FT253

Nguyen, D., Dogruöz, S., Rosé, C., and de Jong, F. (2016). Computational sociolinguistics: a survey. *Comput. Ling.* 42, 537–593. doi: 10.1162/COLI_a_00258

Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., and Pentland, A. (2013). Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* 4:1961. doi: 10.1038/ncomms2961

Perez, C., and Germon, R. (2016). Graph creation and analysis for linking actors: application to social data. *Autom. Open Source Intell.* 7, 103–129. doi: 10.1016/B978-0-12-802916-9.00007-5

Preisler, B. (2003). English in Danish and the Danes' English. *Int. J. Soc. Lang.* 159, 109–126. doi: 10.1515/ijsl.2003.001

Raumolin-Brunberg, H. (1996). "Social factors and pronominal change in the seventeenth-century: The Civil War effect?" in *Advances in English Historical Linguistics,* eds J. Fisiak and M. Krygier (Berlin: Mouton de Gruyter), 361–388.

Tagliamonte, S. A., and D'Arcy, A. (2009). Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85, 58–108. doi: 10.1353/lan.0.0084

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 15 July 2020
doi: 10.3389/frai.2020.00048

General Northern English. Exploring Regional Variation in the North of England With Machine Learning

Patrycja Strycharczuk[1]*, *Manuel López-Ibáñez*[2], *Georgina Brown*[3] *and Adrian Leemann*[4]

[1] *Department of Linguistics and English Language, University of Manchester, Manchester, United Kingdom,* [2] *Alliance Manchester Business School, University of Manchester, Manchester, United Kingdom,* [3] *Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom,* [4] *Center for the Study of Language and Society, University of Bern, Bern, Switzerland*

In this paper, we present a novel computational approach to the analysis of accent variation. The case study is dialect leveling in the North of England, manifested as reduction of accent variation across the North and emergence of General Northern English (GNE), a pan-regional standard accent associated with middle-class speakers. We investigated this instance of dialect leveling using random forest classification, with audio data from a crowd-sourced corpus of 105 urban, mostly highly-educated speakers from five northern UK cities: Leeds, Liverpool, Manchester, Newcastle upon Tyne, and Sheffield. We trained random forest models to identify individual northern cities from a sample of other northern accents, based on first two formant measurements of full vowel systems. We tested the models using unseen data. We relied on undersampling, bagging (bootstrap aggregation) and leave-one-out cross-validation to address some challenges associated with the data set, such as unbalanced data and relatively small sample size. The accuracy of classification provides us with a measure of relative similarity between different pairs of cities, while calculating conditional feature importance allows us to identify which input features (which vowels and which formants) have the largest influence in the prediction. We do find a considerable degree of leveling, especially between Manchester, Leeds and Sheffield, although some differences persist. The features that contribute to these differences most systematically are typically not the ones discussed in previous dialect descriptions. We propose that the most systematic regional features are also not salient, and as such, they serve as sociolinguistic regional indicators. We supplement the random forest results with a more traditional variationist description of by-city vowel systems, and we use both sources of evidence to inform a description of the vowels of General Northern English.

Keywords: vowels, accent features, Northern English, random forests, feature selection, dialect leveling

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Joshua Waxman,
Yeshiva University, United States
Avi Shmidman,
Bar-Ilan University, Israel

Correspondence:
Patrycja Strycharczuk
patrycja.strycharczuk@
manchester.ac.uk

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *26 March 2020*
Accepted: *08 June 2020*
Published: *15 July 2020*

Citation:
Strycharczuk P, López-Ibáñez M,
Brown G and Leemann A (2020)
General Northern English. Exploring
Regional Variation in the North of
England With Machine Learning.
Front. Artif. Intell. 3:48.
doi: 10.3389/frai.2020.00048

1. INTRODUCTION

Dialect leveling is of central interest to sociolinguistics and dialectology. It is linked to dialect contact, and social mobility, and it is thought to arise through "avoidance or attrition of marked variants" (Trudgill, 1986). Such avoidance may lead to variation and change, in which regional variants are replaced with either standard or pan-regional ones. As such changes occur, regional variation is reduced. In the context of British English, there is robust evidence for leveling-type changes (Kerswill, 2003), and we may therefore ask how much regional variation still remains. Conceptually, this is a straightforward question, but empirically, it is not. In this work, we consider difficulties in quantifying the extent of regional variation in speech, and we propose some new methodological and computational solutions in this respect that rely on crowd-sourcing speech data, and quantifying variation with machine learning.

Our focus is on Northern British English, one of the main dialect groups in the UK. Northern British English can be defined in opposition to Southern British English, i.e., through the presence of linguistic features that are found in the North, but not in the South. These features may be syntactic (e.g., the use of the form "give it me" in Northern English), lexical (e.g., "spelk" as a regional variant for "splinter" in Newcastle), phonological or phonetic. We study phonological and phonetic features, understood as accent-specific realizations of specific vowels. Two features that provide a good demarcation between the North and the South in this respect is the presence of the TRAP-BATH split and the FOOT-STRUT split in the South, but not in the North. Consequently, the BATH vowel is shorter and relatively more front in the North, compared to the South, whereas the STRUT vowel is higher in the North compared to the South. This approach can lead us to consider Northern English to be a cluster of distinct but related varieties, which share a specific realization of BATH and STRUT. However, some linguists use the term "General Northern English" (GNE) or "Standard Northern English" emerging as a more coherent variety spoken by certain speakers across the North, as a result of dialect leveling (Whiteside, 1992; Watt, 2002; Honeybone, 2007; Cardoso et al., 2019). GNE speakers can be expected to display typically northern features, like the northern BATH and STRUT, but not other more narrowly defined northern features. For instance, Watt (2002) notes that traditional Tyneside realization of FACE and GOAT as centering diphthongs are avoided by middle-class Tyneside speakers. These speakers are generally shifting toward a pan-northern monophthongal variant, while Southern-standard-like closed diphthongal realizations are also present. Watt argues that many strongly localized accent features are eroding in Tyneside, under the influence of dialect contact. This, however, interacts with a development of a northern (or more narrowly in this case, north-eastern) identity that constrains dialect leveling such that the developing accents, although leveled, still sound distinctively northern. Tension between avoidance of certain regional features, and willing to signal one's northern identity is also noted by Wells (1982b), who says:

> There are many educated northerners who would not be caught dead doing something so vulgar as to pronounce STRUT words with [ʊ], but who would feel it to be a denial of their identity as northerners to say BATH words with anything other than short [a]. (Wells, 1982b, p.354)

To date, the following types of arguments have been proposed as evidence for General Northern English. One type of evidence is attitudinal, and it is expressed by speakers explicitly classifying their own accent as "northern," as opposed to, for instance, "Geordie" (Newcastle) (Watt, 2002). Another type of evidence is gradual disappearance of certain regional features in favor of pan-regional forms, such as the avoidance of centring diphthongs for FACE and GOAT in Tyneside (Watt, 2002), and diphthongisation of the same vowels in York (Haddican et al., 2013). Thirdly, it has been observed that many northern accents participate in the same sound changes, which makes them more similar to one another. A striking example is GOOSE-fronting, which is affecting multiple varieties of English world-wide, including Northern English accents, such as Bradford (Watt and Tillotson, 2001), York (Haddican et al., 2013), Manchester (Baranowski, 2017) and Carlisle (Jansen, 2019). While all this evidence points toward a degree of linguistic homogenization across the North, we may ask whether General Northern English can be considered a coherent variety, or whether it is still an umbrella term for a group of similar, but distinct accents.

We can phrase the same question in terms of classification: is it empirically justified to use labels such as "General Northern English" to describe the speech of some individuals, as opposed to more specific ones, like "middle-class Manchester English?" If geographically diverse northern speakers sound similar, and are thus difficult to localize within the North, we would take that as evidence for GNE. Implicit in this is the assumption that GNE is a middle-class accent. The issue of class is addressed in Cardoso et al. (2019), who investigate attitudes to accents in employment context, stratifying the sample for region and social class. They draw a distinction between GNE (standard, pan-regional and middle-class) vs. Leeds English (non-standard, regional and working-class). The same distinction can apply to Southern British English varieties, where Standard Southern British English is a non-localized standard, whereas Estuary English is an example of more localized, non-standard, working-class speech. The notion that relatively more standard accents are less regionally diverse is well-established in the dialectology of British English (Wells, 1982a). It is also supported by a long line of variationist work that consistently points to fewer regional features in middle-class speakers, compared to working-class speakers [relevant examples from the North of England, include Baranowski and Turton (2015) on Manchester English and Haddican et al. (2013) on York].

While there are indications of increasing homogeneity of middle class speech across the North of England, systematic evidence to support this intuition is limited. In this work, we investigate putative accent convergence in the North systematically, using an audio corpus of Northern English speech, and by using an explicit computational procedure. Traditional dialectology relies on the notion of accent features, and a comparison can be drawn between different accents by way of establishing that particular features are observed in accent A, as opposed to accent B. The more features are shared between two accents, the greater the similarity. This is a somewhat informal approach that essentially relies on expert intuitions about the relevant features for comparison. Such intuitions are eschewed in neighboring fields of computational linguistics and forensic linguistics, where more holistic approaches have been employed to automated accent recognition. Brown and Wormald (2017) propose a method for classifying accents out of a pre-specified pool, using acoustic information from all phones present in a speech sample. The method is based on a distance measure, computed using mel-frequency cepstral coefficients (MFCCs), and supplied to either a simple correlation calculation or a Support Vector Machine (as demonstrated in Brown, 2016). The models that are used within this process can also be supplied to a hierarchical cluster analysis to reveal the relative degrees of similarity that exists among a set of speakers' accents. Alternatively, we can apply a feature analysis to the speaker-specific accent models to estimate which phonemes are contributing most to distinguishing between different accent varieties. However, the method is only able to identify the relevant phones, but it does not provide an insight into how the specific phones differ between different accents.

In this work, we combine aspects of variationist and computational approaches to studying accent variation. We propose a new method for quantifying similarities between accents, based on random-forests based classification. Similarly to Brown and Wormald (2017), this approach allows to identify the features that are most reliable for distinguishing accents, and it provides a methodological solution for identifying key accent features in an explicit way. Unlike Brown and Wormald (2017), we rely on more traditional acoustic measurements, the first two vowel formants. Our approach has the advantage of being linguistically interpretable: we can not only find the vowel phonemes that distinguish different accents, but we can also describe the difference in linguistically meaningful terms, facilitating comparison with earlier descriptive work (e.g., vowel X is lowered in accent A, compared to accent B). This would not be possible if we used MFCCs, although the trade-off is including fewer phonemes (only vowels), and using less comprehensive acoustic information. The specific research questions for the analysis are:

1. To what extent can individual northern cities be systematically distinguished from the rest, based on vowel formant values?
2. Which vowels are the best predictors for each city?

In addition, we provide an up-to-date description of vowel systems in five cities: Leeds, Liverpool, Manchester, Newcastle upon Tyne and Sheffield, as represented in our speaker sample. The data from 105 speakers reading the same passage. In doing so, we follow the more traditional paradigm of plotting vowel spaces in a two dimensional acoustic space, defined by the first two vowel formants.

1.1. Selected Urban Varieties

The accents we examine represent five urban localities in the North of England: Leeds, Liverpool, Manchester, Newcastle upon Tyne and Sheffield. Their relative location in the UK is presented in **Figure 1**. We chose to focus on urban varieties, because our approach relies on a categorical classification, and the different cities provide a robust way of grouping individual speakers geographically. Another motivation is that urban accents are likely to undergo leveling, due to increased speaker mobility and dialect contact. We selected the specific cities based on their shared characteristics: they are all relatively large urban centers in the North of England. An additional consideration was the availability of a sufficient number of speakers in the corpus we used (see section 2.1 for a description of the corpus).

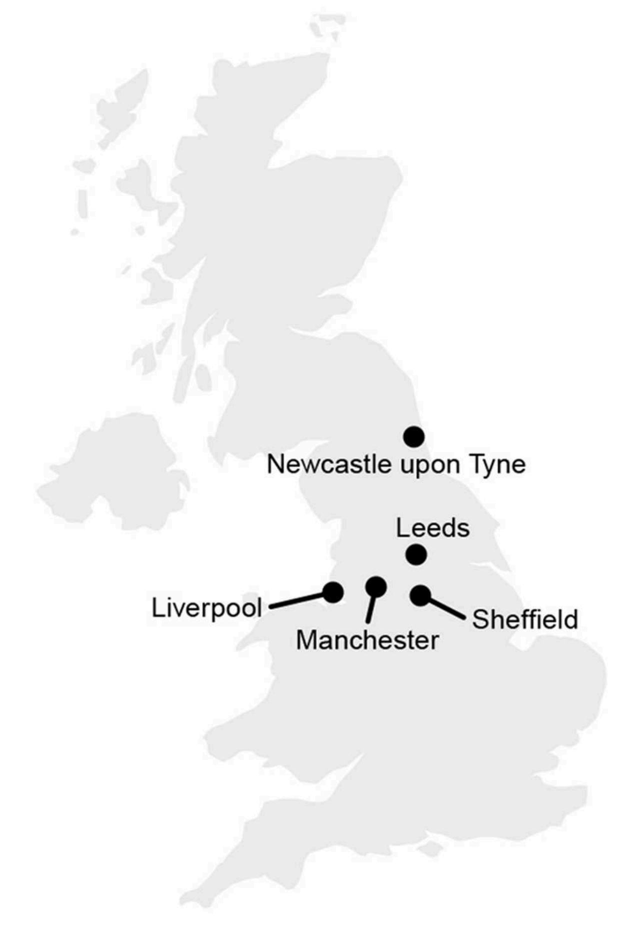


FIGURE 1 | Geographical location of the five selected cities.

In our analysis, we focus on vowels only. This is because we can rely on a well-established method of quantifying differences between vowels, using formant measurements. For consonants, we would have to develop various types of phoneme-specific measurements, and it is less certain that these measurements would capture relevant variation equally well. Additional theoretical motivation for focusing on vowels comes from previous literature which posits that dialect leveling in British English tends to affect vowels more than consonants (Kerswill, 2003).

Below, we provide an overview of previous research on the vowels of the five selected cities. In the description, we use the parameters of variation in describing vowels of English, as developed by Wells (1982c). These are operationalized as lexical sets, selected based on phonemic distinctions in different varieties of English, and abbreviated as keywords. Wells's own description of regional accents are based on auditory transcription within the vowel quadrilateral framework that goes back to Jones (1917). Later works on varieties of English are often supported by acoustic measurements of the first two vowel formants. Recently, comparative dialect studies have been expanded to include articulatory information. We include data from such sources, although we are selective in our literature review, prioritizing sources that include comprehensive vowel descriptions and/or novel observations about specific vowel features. We include some of our own observations about the recordings acquired by the Dynamic Dialect project, where available. The Dynamic Dialects project provides ultrasound and lip video recordings of vowel keywords by speakers of 18 broadly defined accent areas (Lawson et al., 2018). This is a very useful point of reference for readers less familiar with northern accents, as the recordings are recent and freely available online at https://www.dynamicdialects.ac.uk/.

1.1.1. Leeds

Leeds is a city in West Yorkshire, and its accent is described by Wells (1982b) as a prototypically northern. There is no FOOT-STRUT split, or TRAP-BATH split. In addition, according to Wells, Leeds shows some traditional Yorkshire features, such as monophthongisation of FACE, GOAT and PRICE. The realization of SQUARE in Leeds is transcribed as monophthongal by Wells, who also mentions the phonetic quality of NORTH/THOUGHT as being very open. In addition, the happY vowel has a relatively lax quality.

1.1.2. Sheffield

Sheffield is situated in South Yorkshire, and it shares a number of characteristics with Leeds. Among distinct characteristics of the Sheffield dialect, Wells (1982b) mentions a fronted onglide for MOUTH. Additional features of Sheffield English noted by Stoddart et al. (2014) include variable diphthongisation of FLEECE and GOAT, as well as variable fronting of onglide in GOAT. According to Stoddart et al., MOUTH can be monophthongal, and happY is lax.

A more recent description of the Sheffield accent is provided by Williams and Escudero (2014), who compare it to a Southern English system. Their averaged data for Sheffield speakers show diphthongal realization of FACE, GOAT and PRICE, and there is no onglide fronting in MOUTH. These realizations are more similar to Southern English than to the traditional Yorkshire realizations, which is consistent with effects of dialect-leveling. However, statistical comparisons still show differences in the quality of these vowels between Sheffield and the southern accent. The general northern features, absence of [ʌ] and front lax realization of BATH, are apparent in the data. In terms of more recent vowel changes, the 2014 Sheffield English data indicate the presence of GOOSE-fronting, which is, however, less advanced than in the South.

Dynamic Dialects provides ultrasound recordings of two Sheffield speakers. These two speakers vary clearly in their production of FACE and GOAT. One of the speakers produces them as closing diphthongs, whereas the other speaker has more monophthongal variants. The PRICE vowel is diphthongal for both. For both of them, the GOOSE vowel appears somewhat fronted, in line with the data in Williams and Escudero (2014).

1.1.3. Manchester

According to Wells (1982b), Manchester is very similar to Leeds in terms of vowels. However, in an updated description, Baranowski and Turton (2015) stress that FACE and GOAT are closing diphthongs in Manchester, and they do not have a monophthongal quality (this is in contrast to some Lancashire accents). Like other present-day varieties of English, Manchester shows fronting of GOOSE, and to a lesser extent, GOAT. There is no GOAT-fronting before /l/. For GOOSE-fronting, Baranowski and Turton also note an allophonic rule, which is furthermore sensitive to social variation. The GOOSE vowel can be front before /l/ for working-class speakers, but not for middle-class speakers. Similarly, the realization of the STRUT vowel is socially stratified: middle-class speakers show relative lowering of STRUT. SQUARE and NURSE are distinct. Baranowski and Turton (2015) also comments on the realization of happY and lettER vowels. The happY vowel is relatively retracted and lowered, whereas lettER is somewhat retracted. The lettER vowel is reported in some sources to be lowered in Manchester (Beal, 2008). This aspect of Manchester speech is often stereotyped. It is not uncommon to see “Manchester” spelled as “Manchestaaa,” e.g., on social media, as a reference to the quality of the vowel. However, Turton and Ramsammy (2012) observe retraction rather than lowering in lettER.

Data from a single speaker of Manchester English are available through Dynamic Dialects. Interestingly, this speaker has a lowered vowel in STRUT, which is distinct from FOOT, as observed for some Manchester speakers by Turton and Baranowski (2020). This speaker also has diphthongal FACE and GOAT (the onglide of GOAT is also fronted). Her happY vowel is relatively tense. In contrast, she shows the typically northern fronted production of BATH.

1.1.4. Liverpool

Compared to other northern accents, Liverpool is quite distinct, which is attributed to high migration levels into the city from a range of groups (Knowles, 1978). In terms of specific vowel features, Wells (1982b) mentions the merger between SQUARE and NURSE, both of which are realized as a centralized vowel, rounded or unrounded. FACE and GOAT are diphthongal, and there is also a slight diphthongisation of FLEECE and GOOSE. The vowel in happY is tense, unlike in Manchester and Yorkshire. In their study of Liverpool vowels, Ferragne and Pellegrino (2010) confirm this description, and they also note the phonetic proximity of *hod* and *hard* (LOT and START), and between *hid* and *heard* (KIT and NURSE). According to Watson (2007), the PRICE vowel can be monophthongal. Watson also notes optional GOAT-fronting. Furthermore, Cardoso (2015) observes a pattern of phonological variation in PRICE and MOUTH in Liverpool, affecting the trajectories as a function of manner of articulation of the following consonant, and its voicing.

1.1.5. Newcastle Upon Tyne

Traditional Newcastle English shows obvious differences from other northern accents. It is generally reported to display the northern STRUT and FOOT. Wells (1982b) notes that some BATH and TRAP words can have a long [aː], unlike in most other Northern accents. He describes the Newcastle FACE and GOAT vowels as varying between monophthongs and centring diphthongs. FLEECE is said to be “strikingly diphthongal” in final position. The MOUTH vowel is variable, including some traditional [uː] realizations. Among the unstressed vowels, happY is relatively tense, whereas lettER is said to have a particularly open quality.

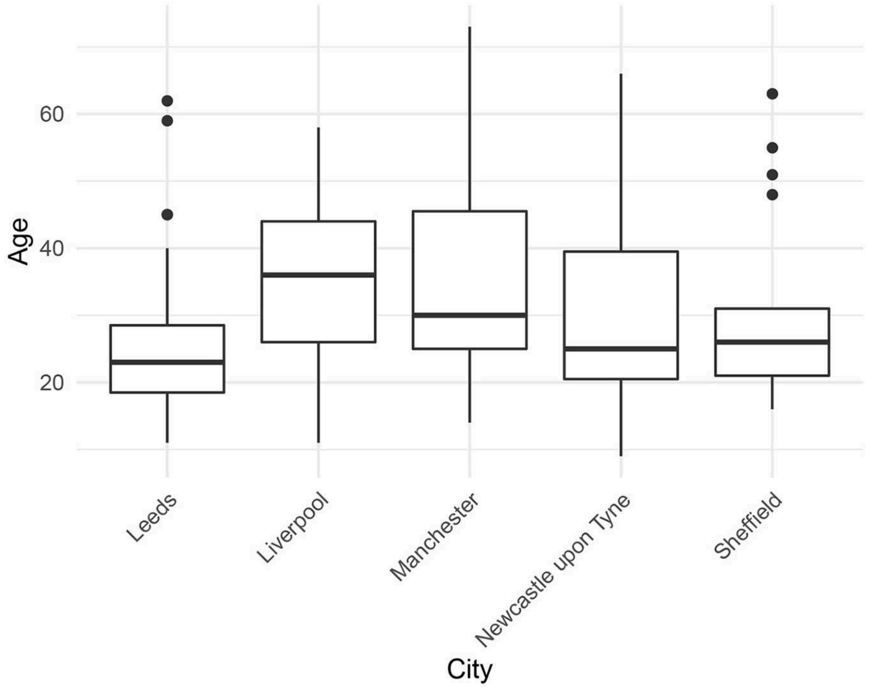


FIGURE 2 | Speaker age by city.

Watt (1998) confirms that FLEECE and GOOSE can be diphthongal in open syllables in Newcastle, whereas closed syllables invariably involve a monophthongal variant. Watt also documents extensively the variation in NURSE, which includes a front rounded variant, as well as a strongly retracted one, and one that is close to Southern British English.

Ferragne and Pellegrino (2010) confirm aspects of this description, adding observations concerning front and close realization of NURSE in Newcastle. They also comment on the variation in FACE and GOAT, including the monophthongal variants and centring diphthongs. Watt (2002) also includes a closing diphthong as a possible variant for FACE and GOAT, and he notes that such realizations are on the rise.

The Dynamic Dialects Newcastle speaker shows a monophthongal GOAT, and a centring diphthong for FACE, with a relatively lowered onglide. This speaker also shows fronting of the onglide in PRICE. His happY is tense. Lowering in lettER is not evident. NURSE is relatively front, and the lip protrusion is evident in the video data. FLEECE is clearly monophthongal (that is in a closed syllable context).

2. MATERIALS AND METHODS

2.1. Corpus

The data we use were extracted from the English Dialects App Corpus (EDAC, Leemann et al., 2018). The data are crowdsourced recordings of the passage "The Boy who Cried Wolf," collected via a mobile phone app. At the time this paper was written, the corpus contained recordings from 3,500 speakers in the British Isles (including Republic of Ireland). Apart from donating the recording, the speakers identified their own accent by placing a pin on a map. This is an important aspect of the method: we do not use any additional criteria for defining an accent as belonging to a specific city, such as mobility, or family history. The speakers also provided demographic information, including age, gender, ethnicity, and level of education. A detailed description of the corpus is in Leemann et al. (2018).

An advantage of the EDAC corpus is that it uses controlled speech materials, which considerably reduces noise in comparing vowel realizations across different speakers and different groups of speakers. This enables us to work with a relatively smaller sample of speakers, compared to what we would have required if he used spontaneous speech. It also considerably reduces data processing time, obviating the need for manual orthographic transcription.

2.2. Speaker Sample Demographics

We selected 105 speakers from the corpus, representing the five cities: Leeds (N =27), Liverpool (N =17), Manchester (N =23), Newcastle upon Tyne (N =19), and Sheffield (N =19). We chose recordings of sufficient quality, excluding those that were incomplete, had excessive background noise, multiple talkers present, etc. The mean speaker age was 31 years (SD =14). **Figure 2** shows the distribution of speaker age by city. The individual cities are comparable in terms of age, although we note that the Leeds and Sheffield speakers were particularly young.

Fifty-nine percent of our speakers were female. As shown in **Table 1**, the balance of gender was similar across the different

TABLE 1 | Summary of gender by city in our speaker sample.

City	Gender	*N*	%
Leeds	Female	16	59.26
Leeds	Male	11	47.4
Liverpool	Female	9	52.94
Liverpool	Male	8	47.06
Manchester	Female	14	68.7
Manchester	Male	9	39.13
Newcastle upon Tyne	Female	11	57.89
Newcastle upon Tyne	Male	8	42.11
Sheffield	Female	12	63.16
Sheffield	Male	7	36.84

TABLE 2 | Summary of ethnicity by city in our speaker sample.

City	Ethnicity	*N*	%
Leeds	Asian	2	7.41
Leeds	Mixed	1	3.70
Leeds	Other	1	3.70
Leeds	White	23	85.19
Liverpool	Mixed	1	5.88
Liverpool	White	16	94.12
Manchester	Asian	2	8.70
Manchester	Mixed	1	4.35
Manchester	White	20	86.96
Newcastle upon Tyne	Asian	1	5.26
Newcastle upon Tyne	Other	1	5.26
Newcastle upon Tyne	White	17	89.47
Sheffield	Black	1	5.26
Sheffield	Mixed	2	153
Sheffield	White	16	84.21

TABLE 3 | Summary of education by city in our speaker sample.

City	Level of education	*N*	%
Leeds	Higher	16	59.26
Leeds	A-level	4	14.81
Leeds	GSCE	4	14.81
Leeds	Lower than GCSE	1	3.70
Leeds	Under 16	1	3.70
Leeds	None	1	3.70
Liverpool	Higher	12	75.9
Liverpool	A-level	2	11.76
Liverpool	GSCE	1	5.88
Liverpool	Lower than GCSE	1	5.88
Liverpool	Under 16	1	5.88
Manchester	Higher	14	68.7
Manchester	A-level	5	21.74
Manchester	GSCE	1	4.35
Manchester	Lower than GCSE	1	4.35
Manchester	Under 16	1	4.35
Manchester	None	1	4.35
Newcastle upon Tyne	Higher	14	73.68
Newcastle upon Tyne	A-level	3	15.79
Newcastle upon Tyne	GSCE	1	5.26
Newcastle upon Tyne	Under 16	1	5.26
Sheffield	Higher	14	73.68
Sheffield	A-level	1	5.26
Sheffield	GSCE	3	15.79
Sheffield	Under 16	1	5.26

cities. In terms of ethnicity, the speakers were predominantly white (87.6%). 4.77% of speakers were Asian, 4.77% were mixed-race. The sample included one black speaker (from Sheffield), and five who did not identify with any of the ethnicity categories. The proportion of white speakers was roughly equal across the cities. The remaining ethnicities were not well-balanced. The by-city ethnicity data are summarized in **Table 2**.

As far as education is concerned, most speakers in a our sample (66.66%) had a higher education degree (BA or professional/vocational equivalent). 14.2% had been educated up to A-level, whereas 9.5% names GSCE as their highest level of education (this was specified as minimum five GSCEs grade A*–C)[1]. 9.5% of speakers had a lower qualification than that, including those that were under 16. The detailed by-city education data are summarized in **Table 3**. The individual cities are comparable in terms of speaker education, in that ca. 80% in each city had A-levels or a higher degree as their level of education. Education is the best proxy we have for social class, although we know that occupation may be a more reliable predictor (Baranowski and Turton, 2018). Based on the education data alone, we cannot conclude that all the speakers in our corpus are middle-class (in fact, that is almost certainly not the case), but we can expect that the corpus contains a substantial proportion of middle-class speakers.

Summing up the demographic data, a typical speaker in our sample is an urban white woman in her 30s with a university degree. This speaker profile differs noticeably from the Non-Mobile Old Rural Male archetype traditionally associated with the dialectological paradigm. However, for the purpose of researching GNE, the sample is well-suited, especially in its education characteristics, as we can expect speakers with higher levels of education to display more standard features and fewer regional ones.

2.3. Materials

As previously mentioned in section 2.1, the speakers read the story of "The Boy Who Cried Wolf." This is a very short text (216 words), which nonetheless contains all English vowels (based on standard descriptions), and so it is appropriate material for investigating English vowels, according to Deterding (1997). We selected one word representing each keyword, as listed in **Table 4**. In selecting the words, we tried to choose monosyllabic words,

[1] GSCE stands for General Certificate of Secondary Education. It is awarded based on individual subject exams generally taken at age 16. An A(Advanced)-level is a further qualification, also awarded based on subject-specific exam results. This qualification is not obligatory, and it also serves as University entry exam.

TABLE 4 | Words selected for measurement with corresponding keywords.

Item	Keyword
feast	FLEECE
fist	KIT
zoo	GOOSE
plan	TRAP
afternoon	BATH
dark	START
thought	THOUGHT
hot	LOT
foot	FOOT
duck	STRUT
third	NURSE
shepherd	DRESS
fool	FOOL
short	NORTH
safety	happY
safety	FACE
homes	GOAT
shouting	MOUTH
time	PRICE
boy	CHOICE
fear	NEAR
air	SQUARE
however	lettER

but it was not always possible. We could not find consistent selection criteria in terms of segmental and prosodic context, so the set is not well-controlled for in that regard. We keep those limitations in mind when analyzing the results. We acknowledge that we could potentially observe more regional variation related to allophonic alternations if we could vary the segmental and prosodic context systematically. All the keywords, bar one, are based on Wells (1982c). As an additional keyword, we included FOOL. This keyword was chosen to capture the fact that for most younger speakers across many varieties of English, a back [uː] vowel can only occur before a coda /l/ (as in *fool*), whereas in other contexts, the GOOSE vowel is fronted to [ʉː] or [y] (Strycharczuk and Scobbie, 2017a). Furthermore, this allophonic variation is sensitive to regional and social variation, such that /uː/-fronting before an /l/ is attested for some speakers in Manchester (Baranowski and Turton, 2015) and Liverpool (Hughes et al., 2012).

2.4. Data Processing

The selected recordings were forced-aligned using an HTK-based forced aligner developed in house. The vowel boundaries were then manually checked by two Undergraduate Research Assistants for all the selected items, listed in **Table 4**. We measured the first two formants automatically, using Praat. For monophthongs, we measured the formants at midpoint. For diphthongs, we used the onglide and offglide as selected time points, defined as 20 and 80% of the vowel duration respectively. The monophthong-diphthong distinction can differ across different accents. We considered all Standard Southern British English (SSBE) diphthongs as potential diphthongs, and measured them at two points, i.e., CHOICE, FACE, GOAT, MOUTH, NEAR, PRICE, and SQUARE. This is based purely on convention, and it should not be taken as a statement about the dynamic characteristic of any vowel. The convention is not perfect. For instance, SQUARE is often monophthongal, whereas FLEECE and GOOSE can be diphthongized. However, making principled decisions about the classification of each vowel in dynamic terms would require a separate in-depth analysis, and as such, it is beyond the scope of our investigation. Our primary interest is in comparing vowels across different accents, and we assume that measuring vowels at consistent time points for different accents should be sufficient to pick out the relevant cross-accent differences in vowel quality.

We used the Linear Predictive Coding algorithm in Praat to extract the measurements, based on 5 formants, 25 ms Gaussian window and 50 Hz pre-emphasis. For male speakers, the maximum formant was set at 5 kHz, whereas for females speakers, it was 5.5 kHz. All the measurements were checked by PS, and hand-corrected wherever tracking errors were spotted. Although manual corrections affect the reproducibility of our measurements, they were deemed necessary, because we rely on one vowel measurement per keyword per speaker, which makes the analysis sensitive to outliers. Ca. 10% of the measurements were hand-corrected.

2.5. Analysis

The formant data were *z*-scored within speaker (a modification of Lobanov, 1971). We used the normalized vowel formant measurements as the input to the random-forest based classification. The purpose of the analysis was to establish how individual urban accents differ from the ones representing other cities. This allows us to assess the distinctness of each accent, and to identify the specific vowel features that set individual northern cities apart. Accuracy of the models was evaluated using leave-one-out cross-validation. We illustrate the procedure using Manchester as an example. For each speaker, we constructed a training dataset by removing this speaker from the data. We then created a bootstrapped sample, with equal number of Manchester and non-Manchester speakers, using the remaining data. We under-sampled the majority class to create a balanced sample. We trained a random forest model on this dataset and tested its accuracy by predicting whether the left-out speaker was from Manchester or not. This procedure was repeated 100 times per speaker, resampling the bootstrapped sample each time, and averaging the predictions, a procedure known as bootstrapped aggregation (bagging, Breiman, 1996). We used the default settings of the current version (1.3–3) of the `party` package, which are the settings suggested for the construction of unbiased conditional random forests by Strobl et al. (2007). In particular, we used `mtry=5`, where `mtry` is defined as the "number of input variables randomly sampled as candidates at each node" (Hothorn et al., 2020), and `minicriterion=0`, where `minicriterion` is a parameter that controls the depth of the trees (`minicriterion=0` grows trees of maximal depth). We

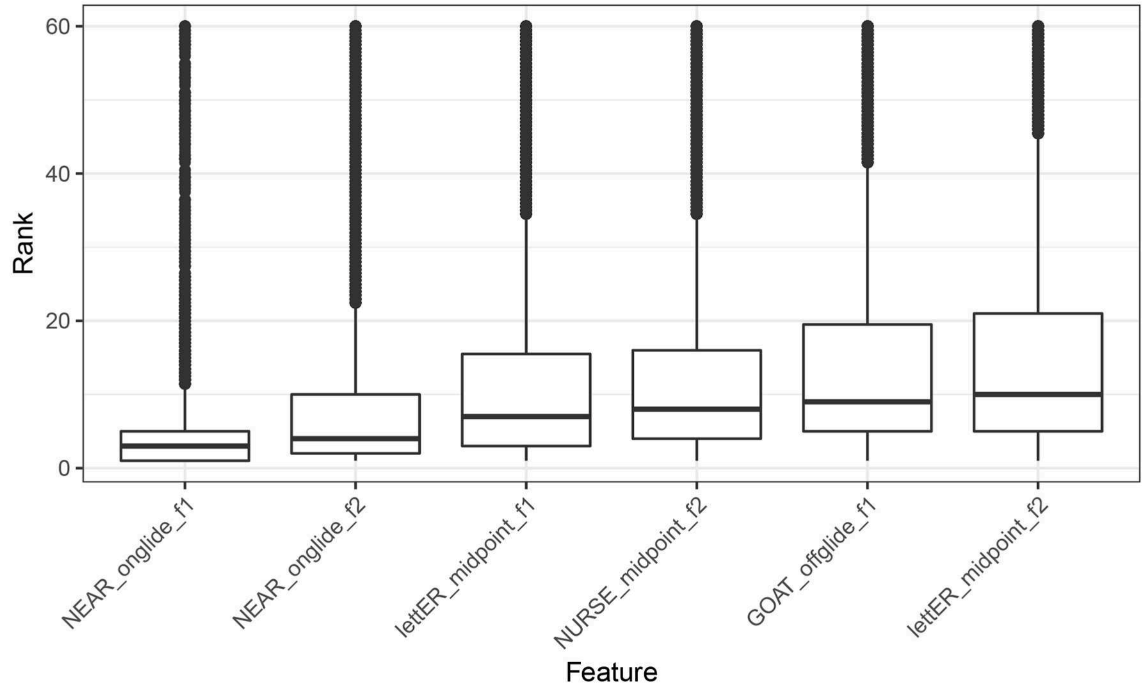


FIGURE 3 | Distribution of feature ranking across all the models for Manchester.

further tested different settings for `mtry`, checking for potential improvements in accuracy, depending on the settings. We find no overall improvement in accuracy for higher values of `mtry` beyond 5.

From each model, we extracted conditional variable importance (Strobl et al., 2008), and ranked the features, according to their relative contribution. We then analyzed the distribution of feature ranking visually, in order to determine which vowels are most consistently used to identify Manchester. The distributions of top ten highest ranking features for Manchester are visualized in **Figure 3**. As we can see, the F1 of NEAR, measured at onglide ranks the highest, followed by onglide F2 for the same vowel. Note that "F1" and "F2" refer to vowel formants here and throughout the paper.

The methodological decisions in setting up the analysis were made to address some of the challenges introduced by the nature of our data. The use of random forests was motivated by the possibility of calculating conditional feature importance, which allows us to identify which input features have the largest influence in the prediction, i.e., which acoustic properties of which vowels set individual cities apart. With the same aim in mind, we set the dependent variable as binary, i.e. Manchester vs. other northern urban accents, Liverpool vs. other northern urban accents, etc. This, however, creates an unbalanced sample, as in each case, the negative category (data from other cities) is about four times bigger than the data from the target city. In order to address this and create balanced data, we used under-sampling. Since under-sampling excludes useful data from the resulting data sample, we used bagging to consider many possible balanced data samples. Given that the data set is relatively small, we were not in a position to split the data into a training test and test set based on a 25–75% split, as is common in random forest modeling. We used leave-one-out cross-validation instead, so we could evaluate accuracy on unseen data, while maximizing the amount of training data.

In order to get more insight into the effect of individual predictors, we used the same bagging process as above (fitting random forest models on a bootstrapped balanced sample that under-samples the majority class), but using only the two features that consistently ranked as most important. We did this without the leave-one-out procedure, and used 1,000 bootstrapped samples per city. We then computed forest predictions for the whole range of values of these features. The output is a heatmap, as visualized in **Figure 4**. The left panel shows the mean, over the 1,000 random forest models, of the probability of predicting Manchester. This is visualized using color scale, where highest certainty of identifying Manchester is associated with relatively darker shade of red. As we can see, the likelihood of an accent being classified as Manchester increases for higher F1 and lower F2 values of the NEAR onglide. Based on established correlations between formant values and tongue height and tongue position, we can interpret this result as follows: Manchester accents are associated with a lowered and centralized onglide for NEAR. The right panel of **Figure 4** shows the standard deviations for the conditional class probabilities.

We also used the formant measurements to generate by-city vowel plots, and we use those for qualitative data analysis. The

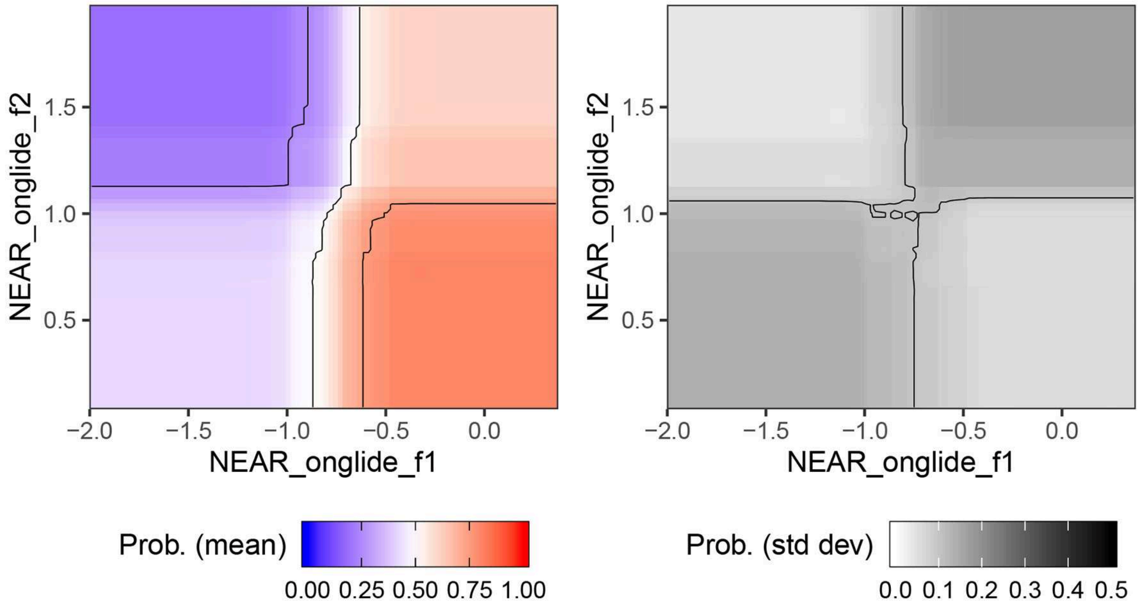


FIGURE 4 | Certainty of the random forests predicting Manchester based on F1 and F2 of NEAR measured at onglide.

vowel plots are in section 3.2, and they show by-vowel median formant values for each city. In order to improve the legibility of the plots, we plot tense monophthongs, lax monophthongs and diphthongs separately. We consider BATH to be lax, based on previous descriptions in northern varieties (see section 1.1). Otherwise, the grouping is based on the same convention as discussed in section 2.4 above.

The data were analyzed in R (R Development Core Team, 2016). The random forests were fitted using the party package 1.3–3 (Hothorn et al., 2006). The vowel plots were generated using modified code originally written by M. Winn (http://www.mattwinn.com/tools.html).

3. RESULTS

3.1. Random-Forest Results

3.1.1. Accuracy

We measure accuracy as the number of correct classifications, using the leave-one-out approach (see section 2.5), as a percentage of the number of trials. **Table 5** provides the accuracy values for each city, along with sensitivity (true positives), and specificity (true negatives) values. Overall, the frequency of correct classification was relatively high for Liverpool (82%) and Newcastle (71%). For the remaining cities, it was lower with 67% for Leeds and 63% for Manchester. For Sheffield, the classification was close to random with 55% accuracy.

If different northern English dialects are becoming more alike, this is predicted to lower the prediction accuracy for the classification models. The overall accuracy results suggest a certain degree of dialect leveling, especially affecting Manchester, Sheffield and Leeds. This is further supported by the accuracy figures broken down for pairs of cities. **Table 6** shows the accuracy figures for each set of forests (forests trained on Leeds as positive category, Manchester as positive category, etc.), in classifying speakers from each remaining city. In this case, correct classification is always negative. This summary confirms that Liverpool and Newcastle are generally well-discriminated from the remaining cities. In contrast, Leeds and Sheffield are highly confusable. Forests trained on Leeds data are more likely than not to classify Sheffield speakers as coming from Leeds. The same situation occurs for models trained on Sheffield: they tend to classify Leeds speakers as coming from Sheffield. There is also a degree of confusability between Leeds and Manchester: classification is close to 50% for this pair of cities, although it is marginally better than random.

TABLE 5 | By-city classification accuracy.

	Leeds	Liverpool	Manchester	Newcastle	Sheffield
Accuracy	67	82	63	71	55
Sensitivity	74	86	72	73	60
Specificity	65	81	60	71	54

3.1.2. Distinguishing Features

The features with the highest median ranking of feature importance for each forest are listed in **Table 8**. The table also provides the direction of prediction for each city, which is based on the heatmaps. We focus on two features for each city, based on the observation that there was typically a large difference in median ranking between the two top features and the rest. This suggests that most forests tend to rely most heavily on the

TABLE 6 | Classification accuracy for pairs of cities.

Predicted city: Leeds (correct if predicts not Leeds)			Predicted city: Sheffield (correct if predicts not Sheffield)			Predicted city: Manchester (correct if predicts not Manchester)		
True city	**% Correct**	**% Incorrect**	**True city**	**% Correct**	**% Incorrect**	**True city**	**% Correct**	**% Incorrect**
Liverpool	93	07	Liverpool	71	29	Newcastle	64	36
Newcastle	73	27	Manchester	57	43	Liverpool	63	37
Manchester	53	47	Newcastle	54	46	Sheffield	62	38
Sheffield	47	53	Leeds	40	60	Leeds	55	45

Predicted city: Liverpool (correct if predicts not Liverpool)			Predicted city: Newcastle upon Tyne (correct if predicts not Newcastle)		
True city	**% Correct**	**% Incorrect**	**True city**	**% Correct**	**% Incorrect**
Leeds	86	14	Liverpool	73	27
Sheffield	83	17	Leeds	75	25
Newcastle	79	21	Manchester	71	29
Manchester	76	24	Sheffield	63	37

TABLE 7 | By-city classification accuracy based on top two features only.

	Leeds	Liverpool	Manchester	Newcastle	Sheffield
Accuracy	64	73	63	65	66
Sensitivity	64	72	64	66	58
Specificity	64	73	63	65	67

same features. We confirmed this by refitting the forests based on top two features for each city only, and analyzing the resulting accuracy. As can be seen in **Table 7**, the accuracy only degrades slightly. We find the biggest drop in accuracy for Liverpool, but the accuracy is nonetheless still high at 73%. For Sheffield, we find an improvement in accuracy, which suggests that having more features leads to overfitting. These results should not be taken to mean that other features do not contribute to the prediction. Since some vowel formants are generally correlated with each other (e.g., THOUGHT and NORTH, diphthongs offglides), we expect that a reasonable degree of accuracy could also be achieved based on different feature combinations, and this is confirmed by exploratory further modeling we have done. Dealing with highly correlated features is one of the strengths of conditional random forests (Strobl et al., 2007), and the known existence of correlations was precisely one of the reasons for choosing this method.

According to forest prediction, the KIT vowel is raised in Leeds, while NORTH is lowered. For Sheffield, the top ranking features are a particularly retracted realization of FOOL and raised onglide of NEAR. The onglide of NEAR is also the most prominent feature for classifying Manchester: in Manchester, the onglide is relatively lowered and centralized.

Based on the random forests, the most systematic features of Liverpool accent are a lowered lettER vowel and a fronted FOOL. Newcastle has a considerably lowered STRUT vowel. The second high-ranking feature for Newcastle is the offglide of PRICE, which is fronted, compared to other cities.

3.2. By-City Vowel Systems

In this section, we present qualitative comparisons of median by-city vowel systems, focusing on the features previously discussed in the literature, summarized in section 1.1. The reader is reminded that the vowel plots are based on one word per vowel, and the segmental context was not controlled for between vowels, but it was consistent between cities (see **Table 4** for the items we used). Therefore, distances between any two vowels within a city might be skewed, but vowels are comparable between cities. Our description is based on medians, and we do not take variance into account at this stage. Therefore, any observed differences should be taken with caution.

We begin with tense monophthongs, illustrated in **Figure 5**. For tense monophthongs, the results seem broadly consistent with previous descriptions. GOOSE is not a back vowel for any of the accents. However, the degree of GOOSE-fronting varies between cities. It is most advanced in Leeds and Manchester, followed by Liverpool, Sheffield and Newcastle. Furthermore, GOOSE is somewhat higher in Leeds, compared to other cities. Furthermore, all cities show considerably more fronting in GOOSE than FOOL. However, there are regional differences in the degree of FOOL-fronting. In Sheffield and Leeds, FOOL is back. The similarity between Sheffield and Leeds in this respect may be one of the factors contributing to the confusability between the two cities, seeing how FOOL retraction is one of the main features of Sheffield. In Liverpool, there is a considerable degree of FOOL-fronting, consistent with what is identified by the random forest analysis. Manchester and Newcastle show in-between median degrees of FOOL-fronting, but the vowel can still be considered back. Another vowel showing some variation is NURSE. It is considerably lowered in Liverpool, compared to other cities. In Newcastle, the median NURSE realization is mid and front-centralized. It resembles most other cities (and SSBE), as opposed to fronted and retracted variants noted in Tyneside (see section 1.1.5). The THOUGHT vowel is somewhat lowered in Leeds. Although the difference is subtle, THOUGHT-lowering is picked out by the random forest analysis as a

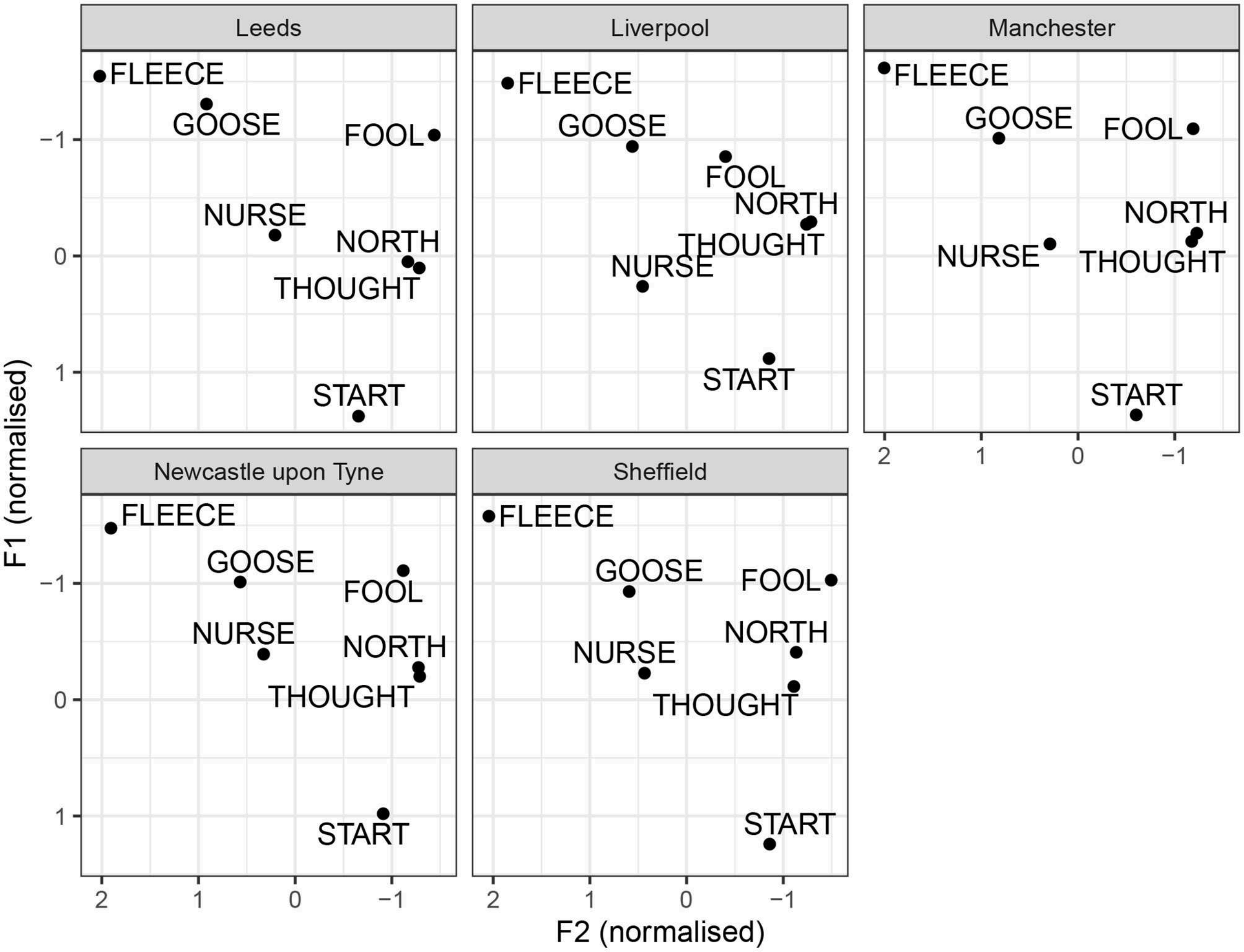


FIGURE 5 | By-city tense vowel systems.

distinguishing feature for Leeds. This is also consistent with Wells's 1982b description of the open quality of THOUGHT in Leeds.

The by-city lax vowel systems are illustrated in **Figure 6**. Once again, these are median values without measures of dispersion. Regional differences can be noted in the F1 of happY. The vowel is higher in Liverpool and Newcastle, compared to Leeds, Sheffield, and Manchester. This is consistent with previous descriptions about the regional distribution of happY-tensing, as present consistently in Liverpool and Newcastle, but not in Manchester or Yorkshire. Nevertheless, happY is higher than KIT for all cities, except Leeds, which is however, due to KIT being exceptionally raised in Leeds (same as in **Table 8**). There does not seem to much evidence for FOOT-fronting in any of the cities, unlike in SSBE (Hawkins and Midgley, 2005; Strycharczuk and Scobbie, 2017b). Note that, in the present data, FOOT tends to have similar degree of acoustic backness to LOT. There seems to be some slight FOOT-fronting in Newcastle, whereas in Liverpool, the FOOT vowel is the most retracted. The STRUT vowel is lower than FOOT for all cities, and it is especially low in Newcastle, where STRUT is clearly distinct from FOOT. The TRAP and BATH vowels show some regional variation in height, but generally BATH is as front as TRAP for all cities. The DRESS vowel is somewhat lowered in Liverpool, compared to other locations. The lettER vowel is very similar in Leeds, Sheffield, and Manchester, but relatively more open in Liverpool and Newcastle.

In comparison to previous descriptions, our results largely confirm that the reports about the regional distribution of happY tensing. They also confirm that, across the North of England, the BATH vowel patterns with TRAP. The lowering of lettER in Newcastle is consistent with the description by Wells (1982b). However, we also find lettER lowering in Liverpool, where it had not been noted. Conversely, the Manchester lettER vowel is not lowered, contra the stereotype. The DRESS vowel also seems lowered, as well as centered in Liverpool. Perhaps most strikingly, for all cities, and especially in Newcastle, we find some STRUT lowering, relative to FOOT.

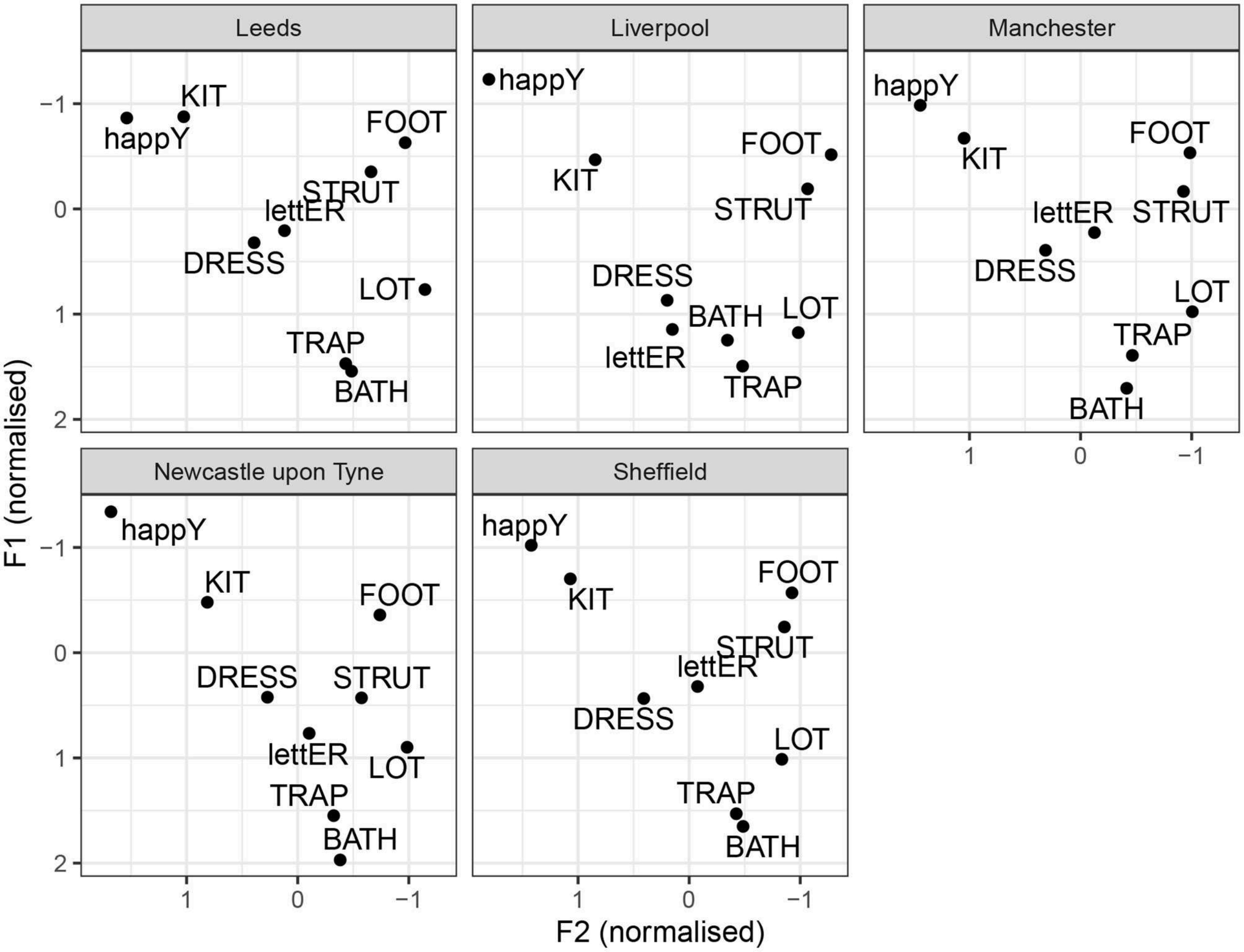


FIGURE 6 | By-city lax vowel systems.

Figure 7 illustrates the diphthongs systems for the individual cities. Impressionistically, the diphthongs appear to show more regional variation than monophthongs. FACE is a closing diphthong for all cities. The median values do not include monophthongal varieties, as reported for Yorkshire, or centring diphthongs, as reported for Newcastle. In Manchester and Liverpool, FACE seems to be more diphthongal, compared to other cities. A similar generalization can be made for GOAT: it is a closing diphthong overall, and it is relatively wider in Manchester and Liverpool. Furthermore, there is regional variation with respect to the onglide of GOAT. In Liverpool, there is quite clear GOAT-fronting. The offglide of GOAT is also more front in Leeds and Sheffield, compared to Manchester and Newcastle. The PRICE vowel has a relatively back and low onglide for all cities. The offglide, however, differs by city. In Liverpool, PRICE is relatively monophthongal, which, however, is likely due to the segmental context, since was followed by a nasal (*time*), and PRICE monophthongisation before nasal is noted for Liverpool by Knowles (1978). In Newcastle, PRICE is a very wide diphthong. The remaining cities have an in-between, but clearly diphthongal variant. The MOUTH vowel is relatively stable across the cities. The NEAR vowel is a centring diphthong in Leeds, Sheffield, and Newcastle. In Liverpool and especially Manchester, it is considerably more monophthongal. In Liverpool, it still seems to have a centring, if a short, trajectory. In Manchester, the offglide is somewhat higher than the onglide, but there is very little movement overall. The SQUARE vowel is quite clearly diphthongal in Newcastle, with a surprisingly low offglide. In comparison, other cities have a more monophthongal variant. In Liverpool, the SQUARE vowel is relatively raised, overlapping in the formant range with NURSE, consistent with previous reports of a NURSE-SQUARE merger.

4. DISCUSSION

The main hypothesis underlying this research is that a large group of speakers in the North of England are converging to a pan-regional standard, and therefore, they cannot be localized further within the North, based on their vowel system. We investigated this by quantifying the success of random forest models trained to differentiate selected Northern English urban accents from a mixed pool of other Northern English accents. Overall, we find that two northern urban accents, Liverpool and Newcastle, remain quite distinct, and therefore they pose few challenges to classification, whereas we do observe a degree of classification uncertainty between Manchester and Leeds accents, and even more so between Leeds and Sheffield accents. From previous descriptions, we would expect that Manchester, Leeds and Sheffield accents are more similar to one another than Newcastle or Leeds. However, our current results allow us to quantify this observation with more precision: while some speakers from these cities can be reliably classified in terms of their accents, in about half of the cases, Leeds and Sheffield speakers in our data are mutually misclassified. Similarly, the classification accuracy for the Manchester—Leeds pair approaches random.

Similarities between the vowel systems of Leeds, Sheffield, and Manchester are further confirmed by the median vowel measurements for each city, as shown in **Figures 5–7**. For example, for all three cities, the typical happY realization is tense,

TABLE 8 | Highest ranked features for each city.

City	Feature	Direction	Articulatory interpretation
Leeds	KIT midpoint F1	lower	vowel is raised
	NORTH midpoint F1	higher	vowel is lowered
Sheffield	FOOL midpoint F2	lower	vowel is retracted
	NEAR onglide F1	lower	onglide is raised
Manchester	NEAR onglide F1	higher	onglide is lowered
	NEAR onglide F2	lower	onglide is retracted
Liverpool	lettER midpoint F1	higher	vowel is lowered
	FOOL midpoint F2	higher	vowel is fronted
Newcastle upon Tyne	STRUT midpoint F1	higher	vowel is lowered
	PRICE offglide F2	higher	offglide is fronted

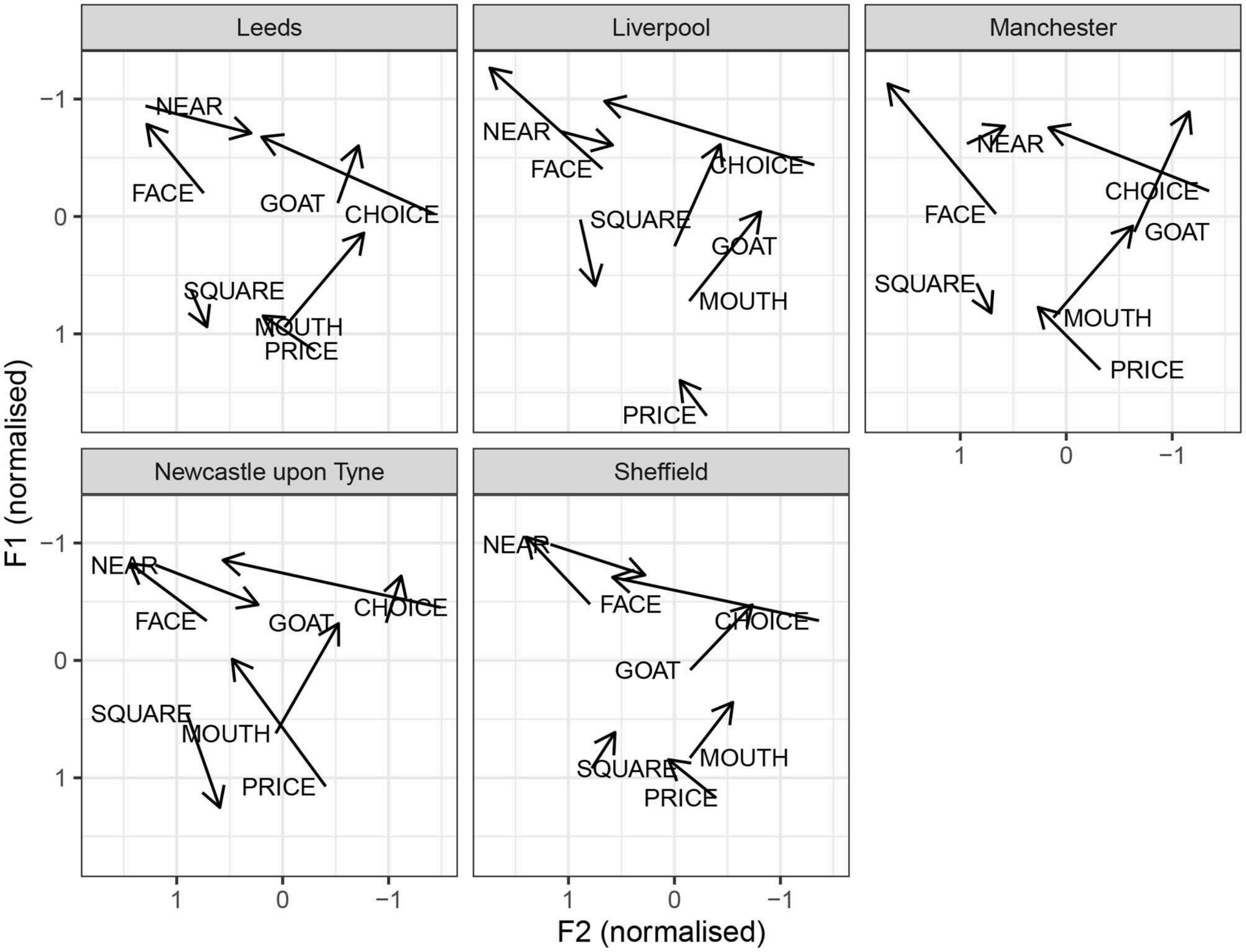


FIGURE 7 | By-city diphthong vowel systems.

GOOSE is fronted, whereas FOOL is retracted, and FACE, PRICE, and CHOICE are all closing diphthongs. All these features are, broadly, also observed in Southern British English, and their robust presence in our data can be taken as a sign of dialect leveling in the North toward a more general British Standard. However, some general northern features prevail, including fronted realization of BATH (consistent with no BATH-TRAP split) and a raised STRUT vowel. The three accents also share a monophthongised realization of SQUARE, which is considered a general northern feature by Honeybone (2007).

The realization of STRUT warrants further comment: while the vowel is relatively raised for all cities (less so for Newcastle), it is not identical to FOOT. This is different from SSBE, but also different from traditional descriptions of Northern English that report no distinction between FOOT and STRUT as one of the identifying features of Northern British English. We must be careful in the phonological interpretation of the phonetic difference we observe. The measurements are not based on a minimal pair, so we cannot be certain that the observed difference in medians is due to a phonemic split between STRUT and FOOT. However, it seems unlikely that a difference of this magnitude would be due to phonetic coarticulation alone. The test items we used were *duck* and *foot*, and there is no reason to expect a strong F1 raising effect in the case of *duck*. We had examined the realization of FOOT and STRUT more systematically in Strycharczuk et al. (2019), using the same corpus, but including more tokens. We found that about 25% of speakers in the corpus have a phonemic split between FOOT and STRUT, while many more have a small but systematic phonetic distinction in the same direction. Thus, the most accurate characterization of the STRUT vowel in the North of England, according to our data, is that the vowel is considerably raised compared to Southern British English [ʌ], but the quality is not necessarily identical to FOOT. A similar observation is made by Turton and Baranowski (2020), based on socially stratified sample of speakers from Manchester. Turton and Baranowski show that the degree of STRUT systematically correlates with social class, with more lowering present in middle-class speakers, compared to working class.

We would argue that the vowel systems for Leeds, Sheffield, and Manchester, as presented in our paper, are all representative of pan-regional General Northern English. At the same time, however, this variety is not a monolith. Some systematic differences between these cities are present in our data. One striking example is the NEAR vowel in Manchester, which has a distinct realization, with a lowered and centralized onglide. Further analysis of sample distribution of F1 and F2 in the onglide of NEAR reveals the presence of even more extremely centralized variants, and these are confined to Manchester. For Leeds, KIT raising is very distinct, and in this case, we see relatively little overlap in F1 values for KIT between Leeds and other cities.

A key outcome of our study is that the features we find to be of most systematic importance in distinguishing individual northern accents are typically not traditional accent features. Among the features listed in **Table 8** only one, NORTH lowering in Leeds is mentioned in a previous description, Wells (1982b), as characteristic of that city. In a way, this is in line with the prediction that dialect-leveling targets salient regional features (Trudgill, 1986; Kerswill, 2003). It is then also expected that less salient regional features may be resistant to leveling. We also believe there is an additional reason why some lesser described features emerge as most important for the classification. To understand this, we need to consider that the success of machine-based classification is facilitated by features that show high-across city and low within-city variation. If the sample from any particular city mostly contains fairly standard speakers, and these speakers make up the most of the training data, the model might not be successful in classifying a speaker who has some very distinct regional features, but who is thereby also very different from the other speakers in the same sample. In contrast, the machine learner performs better with features that are highly consistent, even if the requisite phonetic differences are small. It may also ignore some features that are not consistent within the sample. This is different from a human listener, who is more likely to pay attention to features that are striking, even if such features are less systematic. Translating this distinction into the Labovian paradigm of indicators, markers and stereotypes (Labov, 1972), machine learners will be highly sensitive to indicators, features that systematically distinguish dialects, but that are not the subject of sociolinguistic awareness. It is the absence of sociolinguistic awareness that makes such features systematic within a dialect. Human listeners, on the other hand, are more likely to pick up on markers and stereotypes, by the very definition of markers and stereotypes. This also has consequences in production: speakers are more likely to avoid (some) markers or stereotypes when trying to sound standard.

This point is illustrated by two speakers, each of whom scored 100% accuracy across 100 simulations set up to identify Manchester. This means that 100 models based on different samples, all of which excluded the speaker in question, correctly classified that speaker as coming from Manchester. **Figure 8** shows the formant values for selected vowels, as pronounced by the two speakers. To a linguist, two differences between these two speakers immediately stand out. Speaker 6398 shows has FOOL-fronting, a feature we find in Liverpool English, and which has also been reported in Manchester working class speech. In contrast, speaker 7589 has a retracted FOOL vowel. The two speakers also differ with respect to the FOOT–STRUT contrast: speaker 7589 has a very clear contrast, and the magnitude of the distance seems consistent with a phonological split. Speaker 6398 does not seem to have a difference between FOOT and STRUT, or if there is a difference, it is phonetically marginal. Based on these features, speaker 7589 seems more standard, and in fact, closer to the southern standard, given her pronunciation of STRUT. Speaker 6398, on the other hand, shows clear northern features, including some non-standard ones. However, they both have a lowered and centralized onglide for NEAR. The fact that speaker 7589 incorporates this vowel into an otherwise very standard system corroborates our proposal that lowered and centralized NEAR is an indicator of Manchester speech. This vowel is pronounced differently in Sheffield and Leeds, where the onglide is very close to the offglide of FACE (see **Figure 7**).

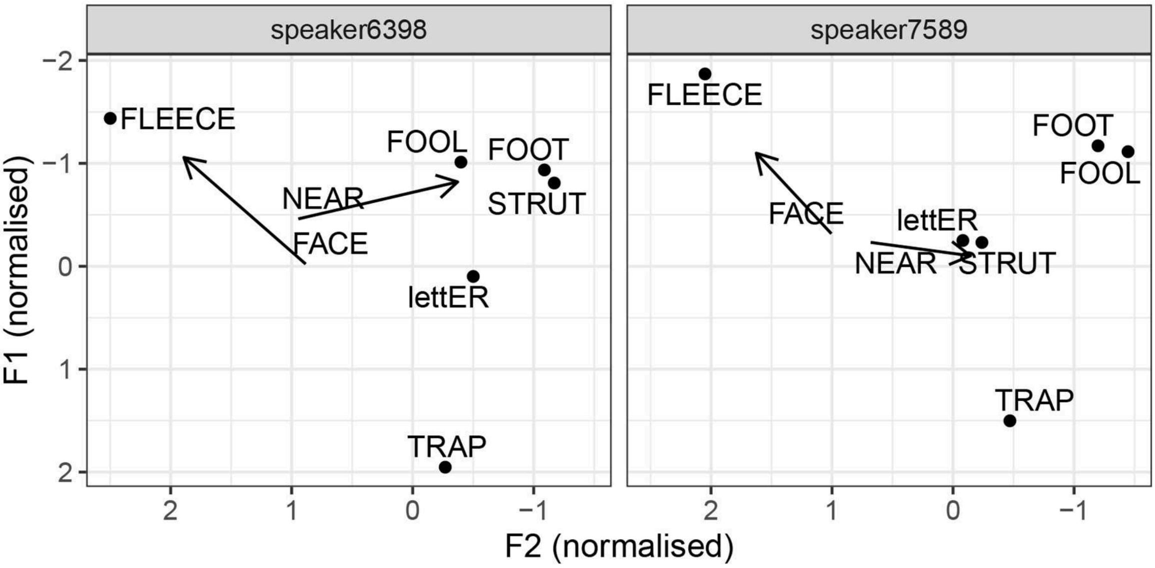


FIGURE 8 | Selected vowels by two Manchester speakers.

The differences between the two speakers in **Figure 8** and the ways they differ from the Manchester median in **Figures 5–7** also bring up an important point about individual variation. We may ask whether these two speakers speak GNE. Are they examples of individual variation within GNE, or do they represent a degree of variation from the standard? If we define GNE as a set of vowel target realizations, then we might be inclined to say the two Manchester speakers are not representative of this variety, or even the Manchester version of it. However, under such a narrow definition, we might find that very few individuals do, in fact, speak GNE. Alternatively, we can also define GNE not by the kind of features we find in the majority of middle-class Northern English speakers, but also by the kind of features we do *not* find. What we do not find is certain marked regional variants, which we can suspect, are perceived by speakers as markers of social class, or lower social prestige. Examples of these include traditional Yorkshire features, such as monophthongised FACE, or a lowered lettER in Manchester. As another possibility, we can define GNE in terms of ranges of possible variation that are set differently for different vowels. For example, there may be a degree of variation possible for the STRUT vowel, such that raised realization as well as some degree of lowering can both be considered GNE. Some regional indicators, as we find in the study, would probably also fit within the permitted range. For instance, the NEAR vowel might be considered standard in both its Manchester and Sheffield variant, even though the two variants clearly differ. Some other vowels, on the other hand, may not vary in the same way. For instance, a speaker with a retracted BATH vowel may be considered standard, but no longer Northern, whereas a speaker with a monopthongised FACE may be seen as northern, but no longer standard.

Liverpool and Newcastle systematically depart from any possible description of GNE. Liverpool accent shows robust local features, including systematic fronting of FOOL, and lowering of lettER. Note that both these features may not be entirely localized to Liverpool, based on previous literature. Sources report FOOL-fronting in working class Manchester speech (Baranowski and Turton, 2015), whereas an open quality of the lettER vowel is, in fact, one of the most stereotyped features of Manchester speech. The fact that, in our data we find these two features to be markers of Liverpool, rather than Manchester, might suggest that the two features carry different social meaning in the two cities. Among other possibilities, they may be more stigmatized in Manchester than in Liverpool, such that more standard Manchester speakers avoid them. Note that Manchester speakers in our sample avoid lettER lowering. If anything, they have a raised lettER vowel compared to other cities. More generally, Liverpool speakers are also likely to differ from other Northern speakers in their attitude toward local features. Although we are not aware of systematic across-city comparisons in this respect, Juskan (2018) presents qualitative data on the attitudes of Liverpool speakers toward their own accent. Some of them explicitly mention the distinctness of Liverpool speech within the UK, and comment on local identity and local pride. A strong sense of local identity is likely to make an accent more resistant to leveling, such that many speakers hold on to at least some regional features. This possibility is consistent with our results. Not only is Liverpool clearly distinct from other cities, but it also shows features that are potentially stigmatized elsewhere in the North (FOOL-fronting). We also find evidence of Liverpool accent preserving its own unique dialect features. For instance, the median vowel formant measurements for NURSE and SQUARE in our data are consistent with there being a NURSE-SQUARE merger in Liverpool, as previously described for this city. Previous research also shows that this feature has relatively low local social prestige (Watson and Clark, 2013), but it resists leveling nonetheless.

Newcastle speech, as represented in our sample, is also distinct, but not because local variants featuring heavily. On the contrary, the Newcastle speakers seem closer to the Southern British standard than the Northern one. One of the salient parameters of variation, in this respect, is that many of the Newcastle speakers had a robust, phonemic-like FOOT-STRUT distinction (this is true for half of the Newcastle speakers in this corpus, as analyzed in Strycharczuk et al., 2019). This finding is similar to the results from Halfacre and Khattab (2019), who report a FOOT-STRUT split in privately educated speakers from Newcastle. The second most prominent feature of Newcastle speech, a fronted offglide of PRICE, is also a feature of standard speech. We also note from **Figure 7** that Newcastle is the only city in our sample with a diphthongal pronunciation of SQUARE, which is also typical of SSBE. Meanwhile, the representative vowel charts do not contain any traditional Newcastle vowel features, such as centring diphthongs in FACE and GOAT. It is not obvious why standard Newcastle speech should be, in a sense, "less northern," than the standard speech of Manchester or Leeds speakers. We can speculate that the social status of the local accent in Newcastle is different than in Manchester or Leeds, such that more standard speakers may avoid blending local features into their speech. Negative attitudes toward traditional accent in Newcastle are mentioned in the context of dialect leveling in this city, as observed by Watt (2002). A related hypothesis is that a raised STRUT vowel is evaluated differently in Newcastle than in other northern UK cities, hence it is not incorporated into the standard. It is also relevant to consider the proximity of Newcastle to the Scottish border. Since Scottish English does not have a FOOT-STRUT split, dialect contact might serve to reinforce the split in neighboring varieties.

Throughout the discussion, we have made references to social meaning in our proposed interpretation of the data. We have set out hypotheses about how specific vowel features may be evaluated, and how such evaluations might differ across the North. Perceptual research is necessary to provide a systematic description of General Northern English. Ultimately, standard speech is defined by what listeners perceive as standard, although it is instructive to see how individuals may deviate from that in production, whether or not consciously. In this context, our research not so much settles all the questions surrounding General Northern English, as it tells us where to look further. Our key contribution is identifying the features that are the loci of systematic regional variation, and features that are not. Further research can determine the relationship between this observed variation and the social perception of standard speech in the North.

In order to identify the features that contribute to differentiate regional accents, we have proposed a novel method, based on random forest classification. This method can be extended to comparing any types of groups that may be of sociolinguistic interest. It can also be extended to include additional features, such as consonantal features, and potentially also to categorical variables. An explicit method for feature selection could be a valuable tool in sociolinguistics, informing researchers' choices of what to study. Currently, the feature choices on the part of sociolinguists are not always overtly motivated. Oftentimes, they are simply the features that researchers notice. However, just like any human listeners, linguists can be biased in their perception, paying special attention to features they know about from previous literature, to features that are marked, and to phonetic differences that are big. One unfortunate outcome of this situation is that instances of small but systematic variation can be systematically missed. The tool we have developed is particularly good at identifying such variation, and as such, it can inform research decisions. Due to its success with identifying regional indicators, the method may have also applications in forensic contexts, such as accent profiling.

We developed the method specifically to maximize the returns from using a relatively small speaker sample. From a computational perspective, our sample ($N = 105$) is indeed small. However, it is a fairly standard number of speakers for a study in speech variation. The practicalities of working with speech seriously limit the amount of data we can presently collect and process. The long-term goal for speech variation studies is to scale up the amount of speech data from different varieties, potentially by pooling different corpora. Such work is already under way (e.g., Stuart-Smith et al., 2020), although we are still some way away from having rich large-scale spoken English corpora with good geographical coverage. In the meantime, trying to mitigate against the limitations of existing resources allows us to continue documenting speech variation, improving the methods as we go along.

5. CONCLUSION

In this study, we used random-forest based classification to quantify the mutual levels of similarity of vowel systems in different accents. Our interest was in evaluating the hypothesis that dialect leveling in middle-class Northern English speakers has led to convergence toward a pan-regional General Northern English. We do find some evidence of such convergence, although some accents cluster in this respect (Manchester, Leeds, Sheffield), whereas others remain more distinct (Liverpool, Newcastle). Our proposed interpretation of this geographical variation relies on regional variance in language attitude, and differences in the perception of local dialect prestige and local pride. Furthermore, while some traditional accent features may be recessive, most speakers in our sample can still be reliably localized to their particular city. This is often cued by less described, but nevertheless systematic vowel features. This finding is consistent with the prediction that dialect-leveling predominantly targets marked regional features. However, it also highlights that we need to re-evaluate the relevant parameters for variation when updating dialect descriptions. Our study contributes a method for doing that, which combines the benefits of computational approaches (an explicit computational procedure) with being phonetically interpretable, which in turn, bridges our findings with more traditional variationist work.

DATA AVAILABILITY STATEMENT

The data and code used in this study are available at https://osf.io/vtp79/.

AUTHOR CONTRIBUTIONS

This study was initiated by PS. The data were extracted and prepared by GB and AL, using bespoke tools designed by GB for forced alignment. PS and ML-I collaborated on the conception of the analysis. The analysis was designed and executed by ML-I. PS wrote the manuscript, which was read, revised and approved by all authors.

FUNDING

We received support from Manchester Q-Step, who funded two Undergraduate Researcher Assistant positions. The open access publications fees are covered by the University of Manchester Library.

ACKNOWLEDGMENTS

We thank the Undergraduates Research Assistants, Enhui Cheng and Chloe Gornall, for help with processing data. The Research Assistants were funded through Manchester Q-Step.

REFERENCES

Baranowski, M. (2017). Class matters: the sociolinguistics of GOOSE and GOAT in Manchester English. *Lang. Variat. Change* 29, 301–339. doi: 10.1017/S0954394517000217

Baranowski, M., and Turton, D. (2015). "Manchester English," in *Researching Northern Englishes*, ed R. Hickey (Amsterdam; Philadelphia, PA: John Benjamins), 293–316. doi: 10.1075/veaw.g55.13bar

Baranowski, M., and Turton, D. (2018). "Locating speakers in the socioeconomic hierarchy: towards the optimal indicators of social class," in *Paper presented at New Ways of Analysing Variation (NWAV) 47* (New York, NY: New York University).

Beal, J. (2008). "English dialects in the North of England: phonology," in *A Handbook of Varieties of English, volume 1: Phonology*, Schneider, W. Edgar, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton (Berlin: Mouton de Gruyter), 113–133.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Brown, G. (2016). "Automatic accent recognition systems and the effects of data on performance," in *Odyssey: The Speaker and Language Recognition Workshop* (Bilbao). doi: 10.21437/Odyssey.2016-14

Brown, G., and Wormald, J. (2017). Automatic sociophonetics: exploring corpora with a forensic accent recognition system. *J. Acous. Soc. Am.* 142, 422–433. doi: 10.1121/1.4991330

Cardoso, A. (2015). *Dialectology, phonology, diachrony: liverpool English realisations of PRICE and MOUTH* (Ph.D. thesis). The University of Edinburgh, Edinburgh, United Kingdom.

Cardoso, A., Levon, E., Sharma, D., Watt, D., and Ye, Y. (2019). "Inter-speaker variation and the evaluation of British English accents in employment contexts," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Melbourne, VIC), 1615–1619.

Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *J. Int. Phonet. Assoc.* 27, 47–55. doi: 10.1017/S0025100300005417

Ferragne, E., and Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *J. Int. Phonet. Assoc.* 40, 1–34. doi: 10.1017/S0025100309990247

Haddican, B., Foulkes, P., Hughes, V., and Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Lang. Variat. Change* 25, 371–403. doi: 10.1017/S0954394513000197

Halfacre, C., and Khattab, G. (2019). "North-south dividers in privately educated speakers: a sociolinguistic study of received pronunciation using the FOOT-STRUT and TRAP-BATH distinctions in the North East and South East of England," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Melbourne, VIC), 2665–2669.

Hawkins, S., and Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *J. Int. Phonet. Assoc.* 35, 183–199. doi: 10.1017/S0025100305002124

Honeybone, P. (2007). "New-dialect formation in nineteenth century Liverpool: a brief history of Scouse," in *The Mersey Sound: Liverpool's Language, People and Places*, eds A. Grant and C. Grey (Open House Press), 106–140.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. (2006). Survival ensembles. *Biostatistics* 7, 355–373. doi: 10.1093/biostatistics/kxj011

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2020). *party: A Laboratory for Recursive Partytioning*. R package version 1.3-4.

Hughes, V., Haddican, B., and Foulkes, P. (2012). "The dynamics of variation and change in Northern English back vowels," in *Paper Presented at New Ways of Analysing Variation (NWAV) 41 Conference* (Bloomington, IN: Indiana University).

Jansen, S. (2019). Change and stability in GOOSE, GOAT and FOOT: back vowel dynamics in Carlisle English. *English Lang. Linguist.* 23, 1–29. doi: 10.1017/S1360674317000065

Jones, D. (1917). *Everyman's English Pronouncing Dictionary, 1st Edn*. London: Dent.

Juskan, M. (2018). *Production and Perception of Local Variants in Liverpool English: Change, Salience, Exemplar Priming*. Berlin: Language Science Press.

Kerswill, P. (2003). "Dialect levelling and geographical diffusion in British English," in *Social Dialectology: in Honour of Peter Trudgill*, eds D. Britain and J. Cheshire (Amsterdam: John Benjamins), 223–243. doi: 10.1075/impact.16.16ker

Knowles, G. (1978). "The nature of phonological variables in Scouse," in *Sociolinguistic Patterns in British English*, ed P. Trudgill (London: Arnold), 80–90.

Labov, W. (1972). *Sociolinguistic Patterns*. Oxford: Blackwell.

Lawson, E., Stuart-Smith, J., Scobbie, J. M., Nakai, S. (2018). *Dynamic Dialects: An Articulatory Web Resource for the Study of Accents*. University of Glasgow. Available online at: https://www.dynamicdialects.ac.uk/ (accessed June 30, 2020).

Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand* 5, 1–17. doi: 10.1016/j.amper.2017.11.001

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *J. Acous. Soc. Am.* 49, 606–608. doi: 10.1121/1.1912396

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Stoddart, J., Upton, C., and Widdowson, J. D. (2014). "Sheffield dialect in the 1990s: revisiting the concept of NORMs," in *Urban voices: Accent Studies in the British Isles*, eds P. Foulkes and G. Docherty (London; New York, NY: Routledge), 72–89.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinform*. 9:307. doi: 10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform*. 8:25. doi: 10.1186/1471-2105-8-25

Strycharczuk, P., Brown, G., Leemann, A., and Britain, D. (2019). "Investigating the FOOT-STRUT distinction in northern Englishes using crowdsourced data," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds

S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Melbourne, VIC), 1337–1341.

Strycharczuk, P., and Scobbie, J. (2017a). Whence the fuzziness? Morphological effects in interacting sound changes in Southern British English. *Lab. Phonol.* 8:7. doi: 10.5334/labphon.24

Strycharczuk, P., and Scobbie, J. M. (2017b). Fronting of Southern British English high-back vowels in articulation and acoustics. *J. Acous. Soc. Am.* 142, 322–331. doi: 10.1121/1.4991010

Stuart-Smith, J., Sonderegger, M., and Mielke, J. (2020). *Speech Across Dialects of English (spade)*: *Large-Scale Digital Analysis of a Spoken Language Across Space and Time*. ESRC Grant ES/R003963/1, NSERC/CRSNG Grant RGPDD 501771-16, SSHRC/CRSH Grant 869-2016-0006, NSF Grant SMA-1730479.

Trudgill, P. (1986). *Dialects in Contact*. Oxford: Blackwell.

Turton, D., and Baranowski, M. (2020). Not quite the same: The social stratification and phonetic conditioning of the FOOT-STRUT vowels in Manchester. *J. Linguist*. 1–39. doi: 10.1017/S0022226720000122

Turton, D., and Ramsammy, M. (2012). "/ɪ, ə/-lowering in Manchest[ʌ]: contextual patterns of gradient and categorical variabilit[Ë]," in *Paper Presented at 20th Manchester Phonology Meeting* (Manchester).

Watson, K. (2007). Liverpool English. *J. Int. Phonet. Assoc*. 37, 351–360. doi: 10.1017/S0025100307003180

Watson, K., and Clark, L. (2013). How salient is the NURSE~SQUARE merger? *English Lang. Linguist*. 17, 297–323. doi: 10.1017/S136067431300004X

Watt, D. (1998). *Variation and change in the vowel system of Tyneside English* (Ph.D. thesis). Newcastle University, Newcastle upon Tyne, United Kingdom.

Watt, D. (2002). 'I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced levelling in the Tyneside vowel system. *J. Sociolinguist*. 6, 44–63. doi: 10.1111/1467-9481.00176

Watt, D., and Tillotson, J. (2001). A spectrographic analysis of vowel fronting in Bradford English. *English World Wide* 22, 269–303. doi: 10.1075/eww.22.2.05wat

Wells, J. (1982a). *Accents of English 1: An Introduction, Vol. 1*. Cambridge: Cambridge University Press.

Wells, J. (1982b). *Accents of English 2: The British Isles, Vol. 2*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511611766

Wells, J. (1982c). *Accents of English, Vol 3*. Cambridge: Cambridge University Press.

Whiteside, S. (1992). "Analysis-resynthesis: modelling selected phonetic segments of a woman speaker with a General Northern accent," in *Proceedings - Institute of Acoustics* (Windermere), Vol. 14, 511–518.

Williams, D., and Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *J. Acous. Soc. Am*. 136, 2751–2761. doi: 10.1121/1.4896471

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/frai.2020.00055

A Framework for the Computational Linguistic Analysis of Dehumanization

Julia Mendelsohn[1]*, *Yulia Tsvetkov*[2] *and Dan Jurafsky*[3]

[1] School of Information, University of Michigan, Ann Arbor, MI, United States, [2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, United States, [3] Department of Linguistics, Stanford University, Stanford, CA, United States

Dehumanization is a pernicious psychological process that often leads to extreme intergroup bias, hate speech, and violence aimed at targeted social groups. Despite these serious consequences and the wealth of available data, dehumanization has not yet been computationally studied on a large scale. Drawing upon social psychology research, we create a computational linguistic framework for analyzing dehumanizing language by identifying linguistic correlates of salient components of dehumanization. We then apply this framework to analyze discussions of LGBTQ people in the *New York Times* from 1986 to 2015. Overall, we find increasingly humanizing descriptions of LGBTQ people over time. However, we find that the label *homosexual* has emerged to be much more strongly associated with dehumanizing attitudes than other labels, such as *gay*. Our proposed techniques highlight processes of linguistic variation and change in discourses surrounding marginalized groups. Furthermore, the ability to analyze dehumanizing language at a large scale has implications for automatically detecting and understanding media bias as well as abusive language online.

Keywords: computational sociolinguistics, dehumanization, lexical variation, language change, media, *New York Times*, LGBTQ

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Wei Lu,
Singapore University of Technology and Design, Singapore
Jonathan Schler,
Holon Institute of Technology, Israel

****Correspondence:***
Julia Mendelsohn
juliame@umich.edu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *03 March 2020*
Accepted: *29 June 2020*
Published: *07 August 2020*

Citation:
Mendelsohn J, Tsvetkov Y and Jurafsky D (2020) A Framework for the Computational Linguistic Analysis of Dehumanization. Front. Artif. Intell. 3:55. doi: 10.3389/frai.2020.00055

1. INTRODUCTION

Despite the American public's increasing acceptance of LGBTQ people and recent legal successes, LGBTQ individuals remain targets of hate and violence (Dinakar et al., 2012; Silva et al., 2016; Gallup, 2019). At the core of this issue is dehumanization, "the act of perceiving or treating people as less than human" (Haslam and Stratemeyer, 2016), a process that heavily contributes to extreme intergroup bias (Haslam, 2006). Language is central to studying this phenomenon; like other forms of bias (Wiebe et al., 2004; Greene and Resnik, 2009; Recasens et al., 2013; Voigt et al., 2017; Breitfeller et al., 2019), dehumanizing attitudes are expressed through subtle linguistic manipulations, even in carefully-edited texts. It is crucial to understand the use of such linguistic signals in mainstream media, as the media's representation of marginalized social groups has far-reaching implications for social acceptance, policy, and safety.

While small-scale studies of dehumanization and media representation of marginalized communities provide valuable insights (e.g., Esses et al., 2013), there exist no known large-scale analyses, likely due to difficulties in quantifying such a subjective and multidimensional psychological process. However, the ability to do large-scale analysis is crucial for understanding how dehumanizing attitudes have evolved over long periods of time. Furthermore, by being able to account for a greater amount of media discourse, large-scale techniques can provide a more complete view of the media environment to which the public is exposed.

Linguistics and computer science offer valuable methods and insights on which large-scale techniques might be developed for the study of dehumanization. By leveraging more information about the contexts in which marginalized groups are discussed, computational linguistic methods enable large-scale study of a complex psychological phenomenon and can even reveal linguistic variations and changes not easily identifiable through qualitative analysis alone.

We develop a computational linguistic framework for analyzing dehumanizing language, with a focus on lexical signals of dehumanization. Social psychologists have identified numerous components of dehumanization, such as *negative evaluations of a target group*, *denial of agency*, *moral disgust*, and *likening members of a target group to non-human entities, such as vermin*. Drawing upon this rich body of literature, we first identify linguistic analogs for these components and propose several computational techniques to measure these linguistic correlates. We then apply this general framework to explore changing representations of LGBTQ groups in the *New York Times* over 30 years and both qualitatively and quantitatively evaluate our techniques within this case study. We additionally use this lens of dehumanization to investigate differences in social meaning between the denotationally-similar labels *gay* and *homosexual*. We focus on a single case study in order to conduct an in-depth analysis of our methodology, but our framework can generalize to study representations of other social groups, which we briefly explore in our discussion.

This paper aims to bridge the gaps between computational modeling, sociolinguistics, and dehumanization research with implications for several disciplines. In addition to enabling large-scale studies of dehumanizing language and media representation of marginalized social groups, these techniques can be built into systems that seek to capture both conscious and unconscious biases in text. Furthermore, this work has implications for improving machines' abilities to automatically detect hate speech and abusive language online, which are typically underpinned by dehumanizing language. Finally, our case study demonstrates that such computational analyses of discussions about marginalized groups can provide unique insights into language variation and change within sensitive sociopolitical contexts, and help us understand how people (and institutions) use language to express their ideologies and attitudes toward certain social groups.

Trigger Warning: this paper contains offensive material that some may find upsetting, especially in ***Table 4*** *and* ***Table 7****.*

2. BACKGROUND

2.1. Dehumanization

Our lexical semantic analysis involves quantifying linguistic correlates of component psychological processes that contribute to dehumanization. Our approaches are informed by social psychology research on dehumanization, which is briefly summarized here. Prior work has identified numerous related processes that comprise dehumanization (Haslam, 2006). One such component is *likening members of the target group to non-human entities*, such as machines or animals (Haslam, 2006; Goff et al., 2008; Kteily et al., 2015). By perceiving members of a target group to be non-human, they are "outside the boundary in which moral values, rules, and considerations of fairness apply" (Opotow, 1990), which thus leads to violence and other forms of abuse. Metaphors and imagery relating target groups to vermin are particularly insidious and played a prominent role in the genocide of Jews in Nazi Germany and Tutsis in Rwanda (Harris and Fiske, 2015). More recently, the vermin metaphor has been invoked by the media to discuss terrorists and political leaders of majority-Muslim countries after September 11 (Steuter and Wills, 2010). According to Tipler and Ruscher (2014), the vermin metaphor is particularly powerful because it conceptualizes the target group as "engaged in threatening behavior, but devoid of thought or emotional desire."

Disgust underlies the dehumanizing nature of these metaphors and is itself another important element of dehumanization. Disgust contributes to members of target groups being perceived as less-than-human and of negative social value (Sherman and Haidt, 2011). It is often evoked (both in real life and experimental settings) through likening a target group to animals. Buckels and Trapnell (2013) find that priming participants to feel disgust facilitates "moral exclusion of out-groups." Experiments by Sherman and Haidt (2011) and Hodson and Costello (2007) similarly find that disgust is a predictor of dehumanizing perceptions of a target group. Both moral disgust toward a particular social group and the invocation of non-human metaphors are facilitated by *essentialist* beliefs about groups, which Haslam (2006) presents as a necessary component of dehumanization. In order to distinguish between human and non-human, dehumanization requires an exaggerated perception of intergroup differences. Essentialist thinking thus contributes to dehumanization by leading to the perception of social groups as categorically distinct, which in turn emphasizes intergroup differences (Haslam, 2006).

According to Haslam (2006), prior work describes "*extremely negative evaluations of others*" as a major component of dehumanization. This is especially pronounced in Bar-Tal's account of delegitimization, which involves using negative characteristics to categorize groups that are "excluded from the realm of acceptable norms and values" (Bar-Tal, 1990). While Bar-Tal's defines delegitimization as a distinct process, he considers dehumanization to be one means of delegitimization. Opotow (1990) also discusses broader processes of moral exclusion, one of which is dehumanization. A closely related process is *psychological distancing*, in which one perceives others to be objects or non-existent (Opotow, 1990). Nussbaum (1999) identifies elements that contribute to the objectification (and thus dehumanization) of women, one of which is *denial of subjectivity*, or the habitual neglect of one's experiences, emotions, and feelings.

Another component of dehumanization is the *denial of agency* to members of the target group (Haslam, 2006). According to Tipler and Ruscher, there are three types of agency: the ability to (1) experience emotion and feel pain (affective mental states), (2) act and produce an effect on their environment (behavioral potential), and (3) think and hold beliefs (cognitive mental states) (Tipler and Ruscher, 2014). Dehumanization typically involves the denial of one or more of these types of agency (Tipler and Ruscher, 2014).

In section 3, we introduce computational linguistic methods to quantify several of these components.

2.2. Related Computational Work

While this is the first known computational work that focuses on dehumanization, we draw upon a growing body of literature at the intersection of natural language processing and social science. We are particularly inspired by the area of automatically detecting subjective language, largely pioneered by Wiebe et al. who developed novel lexical resources and algorithms for this task (Wiebe et al., 2004). These resources have been used as linguistically-informed features in machine learning classification of biased language (Recasens et al., 2013). Other work has expanded this lexicon-based approach to account for the role of syntactic form in identifying the writer's perspective toward different entities (Greene and Resnik, 2009; Rashkin et al., 2016).

These methods have been used and expanded to analyze pernicious, but often implicit social biases (Caliskan et al., 2017). For example, Voigt et al. analyze racial bias in police transcripts by training classifiers with linguistic features informed by politeness theory (Voigt et al., 2017), and Garg et al. investigate historical racial biases through changing word embeddings (Garg et al., 2018). Other studies focus on how people's positions in different syntactic contexts affect power and agency, and relate these concepts to gender bias in movies (Sap et al., 2017) and news articles about the #MeToo movement (Field et al., 2019). There is also a growing focus on identifying subtle manifestations of social biases, such as condescension (Wang and Potts, 2019), microagressions (Breitfeller et al., 2019), and "othering" language (Burnap and Williams, 2016; Alorainy et al., 2019). In addition, our focus on dehumanization is closely related to the detection and analysis of hate speech and abusive language (Schmidt and Wiegand, 2017; ElSherief et al., 2018).

Gender and racial bias have also been identified within widely-deployed NLP systems, for tasks including toxicity detection (Sap et al., 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), coreference resolution (Rudinger et al., 2018), language identification (Blodgett and O'Connor, 2017), and in many other areas (Sun et al., 2019). Given the biases captured, reproduced, and perpetuated in NLP systems, there is a growing interest in mitigating subjective biases (Sun et al., 2019), with approaches including modifying embedding spaces (Bolukbasi et al., 2016; Manzini et al., 2019), augmenting datasets (Zhao et al., 2018), and adapting natural language generation methods to "neutralize" text (Pryzant et al., 2019).

A related line of research has developed computational approaches to investigate language use and variation in media discourse about sociopolitical issues. For example, some work has drawn upon political communication theory to automatically detect an issue's framing (Entman, 1993; Boydstun et al., 2013; Card et al., 2015) through both supervised classification (Boydstun et al., 2014; Baumer et al., 2015) and unsupervised methods, such as topic modeling and lexicon induction (Tsur et al., 2015; Field et al., 2018; Demszky et al., 2019). Scholars have also developed computational methods to identify lexical cues of partisan political speech, political slant in mass media, and polarization in social media (Monroe et al., 2008; Gentzkow and Shapiro, 2010; Demszky et al., 2019).

2.3. Attitudes Toward LGBTQ Communities in the United States

Some background about LGBTQ communities is necessary for our case study of LGBTQ dehumanization in the *New York Times*. Bias against LGBTQ people is longstanding in the United States. Overall, however, the American public has become more accepting of LGBTQ people and supportive of their rights. In 1977, equal percentages of respondents (43%) agreed and disagreed with the statement that gay or lesbian relations between consenting adults should be legal (Gallup, 2019). Approval of gay and lesbian relations then decreased in the 1980s; in 1986, only 32% of respondents believed they should be legal. According to Gallup, attitudes have become increasingly positive since the 1990s, and in 2019, 73% responded that gay or lesbian relations should be legal. The Pew Research center began surveying Americans about their beliefs about same-sex marriage in 2001 and found similar trends (Pew Research Center, 2017). Between 2001 and 2019, support for same-sex marriage jumped from 35 to 61%.

In addition to the public's overall attitudes, it is important to consider the specific words used to refer to LGBTQ people. Because different group labels potentially convey different social meanings, and thus have different relationships with dehumanization, our case study compares two LGBTQ labels: *gay* and *homosexual*. The Gallup survey asked for opinions on legality of "homosexual relations" until 2008, but then changed the wording to "gay and lesbian relations." This was likely because many gay and lesbian people find the word *homosexual* to be outdated and derogatory. According to the LGBTQ media monitoring organization GLAAD, *homosexual*'s offensiveness originates in the word's dehumanizing clinical history, which had falsely suggested that "people attracted to the same sex are somehow diseased or psychologically/emotionally disordered"[1]. Beyond its outdated clinical associations, some argue that the word *homosexual* is more closely associated with sex and all of its negative connotations simply by virtue of containing the word *sex*, while terms, such as *gay* and *lesbian* avoid such connotations (Peters, 2014). Most newspapers, including the *New York Times*, almost exclusively used the word *homosexual* in articles about gay and lesbian people until the late 1980s (Soller, 2018). The *New York Times* began using the word *gay* in non-quoted text in 1987. Many major newspapers began restricting the use of the word *homosexual* in 2006 (Peters, 2014). As of 2013, the *New York Times* has confined the use of *homosexual* to specific references to sexual activity or clinical orientation, in addition to direct quotes and paraphrases[2].

Beyond differences in how LGBTQ people perceive the terms *gay* or *lesbian* relative to *homosexual*, the specific choice of label can affect attitudes toward LGBTQ people. In 2012, Smith et al. (2017) asked survey respondents about either "gay and lesbian rights" or "homosexual rights." Respondents who read the word

[1] https://www.glaad.org/reference/lgbtq

[2] https://www.glaad.org/reference/style

"homosexual" showed less support for LGBTQ rights. This effect was primarily driven by high authoritarians, people who show high sensitivity to intergroup distinctions. The authors posit that *homosexual* makes social group distinctions more blatant than *gay* or *lesbian*. This leads to greater psychological distancing, thus enabling participants to remove LGBTQ people from their realm of moral consideration (Smith et al., 2017). Based on prior research and evolving media guidelines, we expect our computational analysis to show that *homosexual* occurs in more dehumanizing contexts than the label *gay*.

3. OPERATIONALIZING DEHUMANIZATION

In section 2.1, we discussed multiple elements of dehumanization that have been identified in social psychology literature. Here we introduce and quantify lexical correlates to operationalize four of these components: *negative evaluations of a target group*, *denial of agency*, *moral disgust*, and *use of vermin metaphors*.

3.1. Negative Evaluation of a Target Group

One prominent aspect of dehumanization is extremely negative evaluations of members of a target group (Haslam, 2006). Attribution of negative characteristics to members of a target group in order to exclude that group from "the realm of acceptable norms and values" is specifically the key component of *delegitimization*, a process of moral exclusion closely related to dehumanization. We hypothesize that this negative evaluation of a target group can be realized by words and phrases whose connotations have extremely low valence, where valence refers to the dimension of meaning corresponding to positive/negative (or pleasure/displeasure) (Osgood et al., 1957; Mohammad, 2018). Thus, we propose several valence lexicon-based approaches to measure this component: paragraph-level valence analysis, Connotation Frames of perspective, and word embedding neighbor valence. Each technique has different advantages and drawbacks regarding precision and interpretability.

3.1.1. Paragraph-Level Valence Analysis

One dimension of affective meaning is *valence*, which corresponds to an individual's evaluation of an event or concept, ranging from negative/unpleasant to positive/pleasant (Osgood et al., 1957; Russell, 1980). A straightforward lexical approach to measure *negative evaluations of a target group* involves calculating the average valence of words occurring in discussions of the target group. We obtain valence scores for 20,000 words from the NRC VAD lexicon, which contains real-valued scores ranging from zero to one for valence, arousal and dominance. A score of zero represents the lowest valence (most negative emotion) and a score of one is the highest possible valence (most positive emotion) (Mohammad, 2018). Words with the highest valence include *love* and *happy*, while words with the lowest valence include *nightmare* and *shit*.

We use paragraphs as the unit of analysis because a paragraph represents a single coherent idea or theme (Hinds, 1977). This is particularly true for journalistic writing (Shuman, 1894), and studies on rhetoric in journalism often treat paragraphs as the unit of analysis (e.g., Barnhurst and Mutz, 1997; Katajamaki and Koskela, 2006). Furthermore, by looking at a small sample of our data, we found that paragraphs were optimal because full articles often discuss unrelated topics while single sentences do not provide enough context to understand how the newspaper represents the target group. We calculate paragraph-level scores by taking the average valence score over all words in the paragraph that appear (or whose lemmas appear) in the NRC VAD lexicon.

3.1.2. Connotation Frames of Perspective

While paragraph-level valence analysis is straightforward, it is sometimes too coarse because we aim to understand the sentiment *directed toward* the target group, not just nearby in the text. For example, suppose the target group is named "B." A sentence, such as "A violently attacked B" would likely have extremely negative valence, but the writer may not feel negatively toward the victim, "B."

We address this by using Rashkin et al.'s Connotation Frames Lexicon, which contains rich annotations for 900 English verbs (Rashkin et al., 2016). Among other things, for each verb, the Connotation Frames Lexicon provides scores (ranging from −0.87 to 0.8) for the writer's perspective toward the verb's subject and object. In the example above for the verb *attack*, the lexicon lists the writer's perspective toward the subject "A," the attacker, as −0.6 (strongly negative) and the object "B" as 0.23 (weakly positive).

We extract all subject-verb-object tuples containing at least one target group label using the Spacy dependency parser[3]. For each subject and object, we capture the noun and the modifying adjectives, as group labels (such as *gay*) can often take either nominal or adjectival forms. For each tuple, we use the Connotation Frames lexicon to determine the writer's perspective toward the noun phrase containing the group label. We then average perspective scores over all tuples.

3.1.3. Word Embedding Neighbor Valence

While a Connotation Frames approach can be more precise than word-counting valence analysis, it limits us to analyzing SVO triples, which excludes a large portion of the available data about the target groups. This reveals a conundrum: broader context can provide valuable insights into the implicit evaluations of a social group, but we also want to directly probe attitudes toward the group itself.

We address this tension by training vector space models to represent the data, in which each unique word in a large corpus is represented by a vector (embedding) in high-dimensional space. The geometry of the resulting vector space captures many semantic relations between words. Furthermore, prior work has shown that vector space models trained on corpora from different time periods can capture semantic change (Kulkarni et al., 2015; Hamilton et al., 2016). For example, diachronic word embeddings reveal that the word *gay* meant "cheerful" or "dapper" in the early twentieth century, but shifted to its current meaning of sexual orientation by the 1970s. Because word embeddings are created from real-world data, they contain real-world biases. For example, Bolukbasi et al. (2016) demonstrated that gender stereotypes are deeply ingrained in these systems.

[3] spacy.io

Though problematic for the widespread use of these models in computational systems, these revealed biases indicate that word embeddings can actually be used to identify stereotypes about social groups and understand how they change over time (Garg et al., 2018).

This technique can similarly be applied to understand how a social group is negatively evaluated within a large text corpus. If the vector corresponding to a social group label is located in the semantic embedding space near words with clearly negative evaluations, that group is likely negatively evaluated (and possibly dehumanized) in the text.

We first preprocess the data by lowercasing, removing numbers, and removing punctuation. We then use the word2vec skip-gram model to create word embeddings (Mikolov et al., 2013). We use Gensim's default parameters with two exceptions; we train our models for ten iterations in order to ensure that the models converge to the optimal weights and we set the window size to 10 words, as word vectors trained with larger window sizes tend to capture more semantic relationships between words (Levy and Goldberg, 2014)[4]. For our diachronic analysis, we first train word2vec on the entire corpus, and then use the resulting vectors to initialize word2vec models for each year of data in order to encourage coherence and stability across years. After training word2vec, we zero-center and normalize all embeddings to alleviate the hubness problem (Dinu et al., 2014).

We then identify vectors for group labels by taking the centroid of all morphological forms of the label, weighted by frequency. For example, the vector representation for the label *gay* is actually the weighted centroid of the words *gay* and *gays*. This enables us to simultaneously account for adjectival, singular nominal, and plural nominal forms for each social group label with a single vector. Finally, we estimate the valence for each group label by identifying its 500 nearest neighbors via cosine similarity, and calculating the average valence of all neighbors that appear in the NRC VAD Valence Lexicon[5].

We also induce a valence score directly from a group label's vector representation by adapting the regression-based sentiment prediction from Field and Tsvetkov (2019) for word embeddings. This approach yielded similar results as analyzing nearest neighbor valence but was difficult to interpret. More details for and results from this technique can be found in the **Supplementary Material**.

3.2. Denial of Agency

Denial of agency refers to the lack of attributing a target group member with the ability to control their own actions or decisions (Tipler and Ruscher, 2014). Automatically detecting the extent to which a writer attributes cognitive abilities to a target group member is an extraordinarily challenging computational task. Fortunately, the same lexicons used to operationalize *negative evaluations* provide resources for measuring lexical signals of *denial of agency*.

3.2.1. Connotation Frames

As in section 3.1, we use Connotation Frames to quantify the amount of agency attributed to a target group. We use Sap et al.'s extension of Connotation Frames for agency (Sap et al., 2017). Following Sap et al.'s interpretation, entities with high agency exert a high degree of control over their own decisions and are active decision-makers, while entities with low agency are more passive (Sap et al., 2017). This contrast is particularly apparent in example sentences, such as *X searched for Y* and *X waited for Y*, where the verb *searched* gives X high agency and *waited* gives X low agency (Sap et al., 2017). Additionally, Sap et al.'s released lexicon for agency indicates that subjects of verbs such as *attack* and *praise* have high agency, while subjects of *doubts* and *needs* have low agency (Sap et al., 2017).

This lexicon considers the agency attributed to subjects of nearly 2,000 transitive and intransitive verbs. To use this lexicon to quantify *denial of agency*, we extract all sentences' head verbs and their subjects, where the subject noun phrase contains a target group label. Unlike Rashkin et al.'s real-valued Connotation Frames lexicon for perspective, the agency lexicon only provides binary labels, so we calculate the fraction of subject-verb pairs where the subject has high agency.

3.2.2. Word Embedding Neighbor Dominance

The NRC VAD Dominance Lexicon provides another resource for quantifying dehumanization (Mohammad, 2018). The NRC VAD lexicon's dominance dimension contains real-valued scores between zero and one for 20,000 English words. However, the dominance lexicon primarily captures power, which is distinct from but closely related to agency. While power refers to one's control over others, agency refers to one's control over oneself. While this lexicon is a proxy, it qualitatively appears to capture signals of *denial of agency*; the highest dominance words are *powerful*, *leadership*, *success*, and *govern*, while the lowest dominance words are *weak*, *frail*, *empty*, and *penniless*. We thus take the same approach as in section 3.1.3, but instead calculate the average dominance of the 500 nearest neighbors to each group label representation[5].

As in section 3.1.3, we also induced a dominance score directly from a group label's vector representation by adapting the regression-based sentiment prediction from Field and Tsvetkov (2019) for word embeddings. More details and results for this technique can be found in the **Supplementary Material**.

3.3. Moral Disgust

To operationalize *moral disgust* with lexical techniques, we draw inspiration from Moral Foundations theory, which postulates that there are five dimensions of moral intuitions: care, fairness/proportionality, loyalty/ingroup, authority/respect, and sanctity/purity (Haidt and Graham, 2007). The negative end of the sanctity/purity dimension corresponds to moral disgust. While we do not directly

[4] https://radimrehurek.com/gensim/models/word2vec.html

[5] We conducted additional analyses by considering 25, 50, 100, 250, and 1,000 nearest neighbors, which yielded similar results and can be found in the **Supplementary Material**.

TABLE 1 | Overview of linguistic correlates and our operationalizations for four elements of dehumanization.

Dehumanization element	Operationalization
Negative evaluation of target group	Paragraph-level sentiment analysis Connotation frames of perspective Word embedding neighbor valence
Denial of agency	Connotation frames of agency Word embedding neighbor dominance
Moral disgust	Vector similarity to disgust
Vermin metaphor	Vector similarity to vermin

incorporate Moral Foundations Theory in our framework for dehumanization, we utilize lexicons created by Graham et al. (2009) corresponding to each moral foundation. The dictionary for moral disgust includes over thirty words and stems, including *disgust**, *sin*, *pervert*, and *obscen** (the asterisks indicate that the dictionary includes all words containing the preceding prefix)[6].

We opt for a vector approach instead of counting raw frequencies of moral disgust-related words because such words are rare in our news corpus. Furthermore, vectors capture associations with the group label itself, while word counts would not directly capture such associations. Using the word embeddings from section 3.1.3, we construct a vector to represent the *concept* of moral disgust by averaging the vectors for all words in the "Moral Disgust" dictionary, weighted by frequency. This method of creating a vector from the Moral Foundations dictionary resembles that used by Garten et al. (2016). We identify implicit associations between a social group and moral disgust by calculating cosine similarity between the group label's vector and the Moral Disgust concept vector, where a higher similarity suggests closer associations between the social group and moral disgust.

3.4. Vermin as a Dehumanizing Metaphor

Metaphors comparing humans to vermin have been especially prominent in dehumanizing groups throughout history (Haslam, 2006; Steuter and Wills, 2010). Even if a marginalized social group is not directly equated to vermin in the press, this metaphor may be invoked in more subtle ways, such as through the use of verbs that are also associated with vermin (like *scurry* as opposed to the more neutral *hurry*) (Marshall and Shapiro, 2018). While there is some natural language processing work on the complex task of metaphor detection (e.g., Tsvetkov et al., 2014), these systems cannot easily quantify such indirect associations.

We thus quantify the metaphorical relationship between a social group and vermin by calculating similarities between these concepts in a distributional semantic vector space. As with *moral disgust*, we create a *Vermin* concept vector by averaging the following vermin words' vectors, weighted by frequency: *vermin*, *rodent(s)*, *rat(s) mice*, *cockroaches*, *termite(s)*, *bedbug(s)*, *fleas*[7]. We do not include the singular *mouse* or *flea* because non-vermin senses of those words were more frequent, and word2vec does not account for polysemy. We calculate cosine similarity between each group label and the *Vermin* concept vector, where a high cosine similarity suggests that the group is closely associated with vermin.

Table 1 provides an overview of the four elements of dehumanization that we study and the lexical techniques used to quantify them.

4. DATA

The data for our case study spans over 30 years of articles from the *New York Times*, from January 1986 to December 2015, and was originally collected by Fast and Horvitz (2016). The articles come from all sections of the newspaper, such as "World," "New York & Region," "Opinion," "Style," and "Sports." Our distributional semantic methods rely on all of the available data in order to obtain the most fine-grained understanding of the relationships between words possible. For the other techniques, we extract paragraphs containing any of the following words from a predetermined list of **LGTBQ terms**: *gay(s), lesbian(s), bisexual(s), homosexual(s), transgender(s), transsexual(s), transexual(s), transvestite(s), transgendered, asexual, agender, aromantic, lgb, lgbt, lgbtq, lgbtqia, glbt, lgbtqqia, genderqueer, genderfluid, intersex, pansexual.*

Each acronym label is matched insensitive to case and punctuation. Some currently prominent LGBTQ terms, such as *queer* and *trans* are not included in this study, as other senses of these words were more frequent in earlier years. We filter out paragraphs from sections that typically do not pertain to news, such as "Arts," "Theater," and "Movies." While these sections could provide valuable information, we focus on representation of LGBTQ groups in more news-related contexts.

A challenging question when analyzing mass media for subjective attitudes is deciding whose perspective we want to capture: an individual reporter, the institution, or society at large? In this case study, we aim to identify the institution's dehumanizing attitudes toward LGBTQ people. We represent the *New York Times* institution as a combination of the journalists' words in news articles, direct quotes, paraphrases from interviews, and published opinion articles. Therefore, despite our news focus, we include data from "Opinion" sections; while opinion articles are stylistically different from traditional journalistic reporting due to more overt biases and arguments, these articles are important in constructing the institution's perspective. In addition, we consider all text in each relevant paragraph, including quotes and paraphrases, because they are important to a newspaper's framing of an issue, as particular quotes representing specific stances are intentionally included or excluded from any given article (Niculae et al., 2015).

We refer to the remaining subset of the *New York Times* data after filtering as the *LGBTQ corpus*. The *LGBTQ* corpus consists of 93,977 paragraphs and 7.36 million tokens. A large increase in reporting on LGBTQ-related issues has led to a skewed distribution in the amount of data over years, with

[6] https://www.moralfoundations.org/othermaterials

[7] Largely inspired by https://en.wikipedia.org/wiki/Vermin

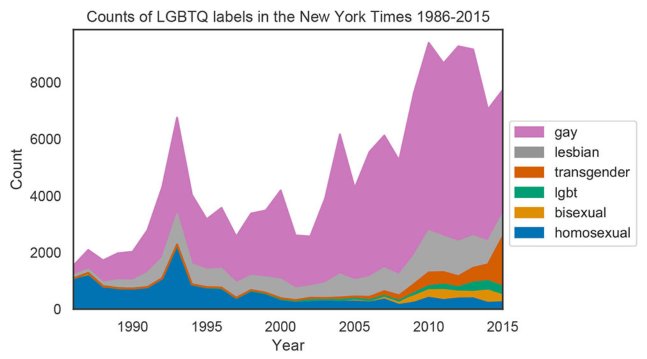


FIGURE 1 | Counts for the six most frequent LGBTQ labels in each year of the *New York Times* data.

TABLE 2 | Nearest words to weighted average of all LGBTQ terms' vectors in 1986, 2000, and 2015.

1986	2000	2015
Sex	Interracial	Sex
Premarital	Openly	Non-transgender
Sexual	Unwed	Unmarried
Abortion	Homophobia	Interracial
Promiscuity	Premarital	Closeted
Polygamy	Ordination	Equality
Promiscuous	Non-whites	Couples
Vigilantism	Ordaining	Abortion
Bestiality	Discrimination	Sexuality
Pornography	Abortion	Antiabortion

1986 containing the least data (1,144 paragraphs and 73,549 tokens) and 2012 containing the most (5,924 paragraphs and 465,254 tokens).

For all experiments, we also include results for the terms *American* and *Americans*. We include *American(s)* to contrast changes in LGBTQ labels' representation with another social group label. This ensures that the changes we find in dehumanizing language toward LGBTQ groups do not apply uniformly to all social groups, and are thus not merely an artifact of the publication's overall language change. While a natural "control" variable would be labels, such as *straight* or *heterosexual*, these terms only occurred within discussions of LGBTQ communities because they name socially unmarked categories. We also considered comparing LGBTQ labels to *person/people*, but because word embedding-based experiments are sensitive to syntactic forms, we opt for a label that behaves more syntactically similar to *gay* and *homosexual*, particularly with both nominal and adjectival uses. Nevertheless, *American(s)* is by no means a neutral control variable. Because of its in-group status for the *New York Times* (a U.S. institution), we expect our measurements to show that *American(s)* appears in more humanizing contexts than LGBTQ labels; however, we do not expect to find substantial changes in the use of *American(s)* over time.

Figure 1 shows the counts of group labels for each year in the *New York Times* from 1986 to 2015. For visualization purposes, only words with a total count >1,000 are shown. The relative frequency of *homosexual* decreased substantially over time, while *gay*, *lesbian*, and *bisexual* are more frequent in later years. The terms *lgbt* and *transgender* also emerged after 2000. Counts for all LGBTQ labels can be found in the **Supplementary Material**.

5. RESULTS

5.1. Word Embeddings

Using all of the *New York Times* data, we create word2vec models for each year using the methods described in section 3.1.3. Because our computational techniques rely upon these word2vec models, it is useful to gain a sense of how LGBTQ terms are semantically represented within these models. We thus inspect the ten nearest neighbors, or most similar words, to LGBTQ terms in different years. Note that the neighboring words in **Tables 2**, **3** are shown purely for qualitative investigation; our measures for quantifying each dehumanization component incorporate far more information from the word2vec models beyond the top ten neighbors.

Table 2 shows the 10 nearest neighbors (by cosine similarity) to our vector representation of all LGBTQ terms, which is the weighted average of the embeddings of all LGBTQ terms considered. For visual convenience, we filter out words occurring fewer than ten times, proper names, as well as other LGBTQ labels and forms of the word *heterosexual*, which are common neighbors for all terms across all years.

Table 2 shows that in 1986, LGBTQ groups were most highly associated with words that often convey a sense of sexual deviancy, including *promiscuity*, *promiscuous*, *polygamy*, *bestiality*, and *pornography*. These associations suggest that LGBTQ people were dehumanized to some extent at this time, and their identities were not fully recognized or valued. This shifted by 2000, where we no longer see associations between LGBTQ groups and ideas that evoke moral disgust. Instead, the 2000 vector space shows that LGBTQ people have become more associated with civil rights issues (suggested by *interracial*, *homophobia*, and *discrimination*). The words *ordination* and *ordaining* likely appear due to major controversies that arose at this time about whether LGBTQ people should be permitted to be ordained. We also see some indications of self-identification with the term *openly*. Finally, we see a slight shift toward associations with identity in 2015, with nearby words including *nontransgender*, *closeted*, *equality*, and *sexuality*. Curiously, the word *abortion* is a nearby term for all 3 years. Perhaps this is because opinions toward abortion and LGBTQ rights seem to be divided along similar partisan lines.

Table 3 shows the ten nearest neighboring words to our representations of *gay* and *homosexual* after filtering out proper names, words appearing <10 times that year, other LGBTQ terms, and forms of *heterosexual*. **Table 3** reveals variation in social meaning between *gay* and *homosexual* despite denotational similarity, and these differences intensify over time. In 1986,

TABLE 3 | Nearest words to vector representations of *gay* and *homosexual* in 1986, 2000, and 2015.

1986		2000		2015	
Gay	**Homosexual**	**Gay**	**Homosexual**	**Gay**	**Homosexual**
Homophobia	Premarital	Interracial	Premarital	Interracial	Premarital
Women	Abortion	Openly	Openly	Sex	Sexual
Feminist	Sexual	Homophobia	Deviant	Couples	Bestiality
Vigilante	Sex	Unwed	Interracial	Mormons	Pedophilia
Vigilantism	Promiscuity	Ordination	Promiscuity	Marriage	Adultery
Suffrage	Polygamy	Premarital	Immoral	Closeted	Infanticide
Sexism	Anal	Abortion	Sexual	Equality	Abhorrent
A.c.l.u.	Intercourse	Antigay	Criminalizing	Abortion	Sex
Amen	Consenting	Discrimination	Polygamy	Unmarried	Feticide
Queer	Consensual	Marriagelike	Consensual	Openly	Fornication

gay is associated with terms of discrimination, civil rights and activism, such as *homophobia*, *feminist*, *suffrage*, *sexism*, and *a.c.l.u*. On the other hand, *homosexual* is primarily associated with words related to sexual activity (e.g., *promiscuity*, *anal*, *intercourse*, *consenting*).

In 1986, this pattern may be due to discussions about sexual transmission of AIDS, but the pejoration of *homosexual* continues over time. While *gay* becomes associated with issues related to marriage equality and identity in 2015, *homosexual* becomes extremely associated with moral disgust and illicit activity, with nearest neighbors including *bestiality*, *pedophilia*, *adultery*, *infanticide*, and *abhorrent*.

This qualitative analysis of word embedding neighbors reveals significant variation and change in the social meanings associated with LGBTQ group labels, with clear relationships to dehumanizing language. We will now present our quantitative results for measuring each component of dehumanization.

5.2. Negative Evaluation Toward Target Group

5.2.1. Quantitative Results

5.2.1.1. Paragraph-level valence analysis

Figure 2A shows the average valence for paragraphs containing LGBTQ labels [and *American(s)* for comparison], where a paragraph's valence is simply the average valence over its words (or lemmas) that appear in the NRC VAD Valence Lexicon. The NRC VAD lexicons actually contain several LGBTQ terms, which all have lower than the average valence score of 0.5: *transsexual* (0.264), *homosexual* (0.333), *lesbian* (0.385), *gay* (0.388), and *bisexual* (0.438). These values contrast starkly with more positively-valenced entries in the lexicon, such as *heterosexual* (0.561), *person* (0.646), *human* (0.767), *man* (0.688), and *woman* (0.865). These disparities likely reveal biases among the human annotators whose judgments were used to construct the NRC VAD lexicon (Mohammad, 2018). While the lexicon may itself be an interesting artifact of dehumanizing attitudes toward LGBTQ people, we remove these terms before calculating paragraph-level valence scores in order to isolate linguistic signals in the *New York Times* data from annotation biases. Without this preprocessing step, the temporal trends and relative differences between *all LGBTQ terms*, *gay*, and *homosexual* remain roughly the same, but all LGBTQ labels occur in significantly more negative paragraphs than *American*.

Figure 2A shows the average paragraph valence. For visualization purposes, we present the results over 5-year intervals due to data sparsity in later years for *homosexual* (there were just 208 paragraphs containing *homosexual* in 2014, relative to 3,669 containing *gay* in the same year). Analysis of overlapping confidence intervals and Wilcoxon signed-rank tests over the means for each of the 30 years indicates that *gay* and *all LGBTQ terms* occur in significantly more positive paragraphs than *homosexual* ($p < 0.0001$). A linear regression analysis over all years reveals that *all LGBTQ terms*, *gay*, and *homosexual* all significantly increase in paragraph-level valence over time ($p < 0.0001$). However, when considering just the last 15 years, *gay* still significantly increases in paragraph-level valence, while *homosexual* may be trending downward, although this trend does not reach significance in our data ($p = 0.078$).

The paragraph-level valence analysis shown in **Figure 2A** suggests that LGBTQ groups have become increasingly positively evaluated over time, and thus likely less dehumanized in the *New York Times*. However, the slight downward trend in valence for paragraphs containing *homosexual* between 2001 and 2015 suggests that evaluations of people described as *homosexual* have not improved in the same way as those described by other labels.

Finally, this measurement does not support our initial hypothesis that LGBTQ groups have been more negatively evaluated than *American(s)*, but still reveals that the observed trends for LGBTQ labels are not merely artifacts of changing reporting styles, since paragraphs containing *American(s)* show a very different pattern. Overall, this result demonstrates substantial language change in the *New York Times*'s discussion of LGBTQ people as well as variation in the contexts where different group labels appear, particularly *homosexual*.

5.2.1.2. Connotation frames of perspective

Figure 2B shows the writer's average perspective (valence) toward noun phrases containing either any LGBTQ labels, *gay(s)*, *homosexual(s)*, or the comparison group *American(s)* using the

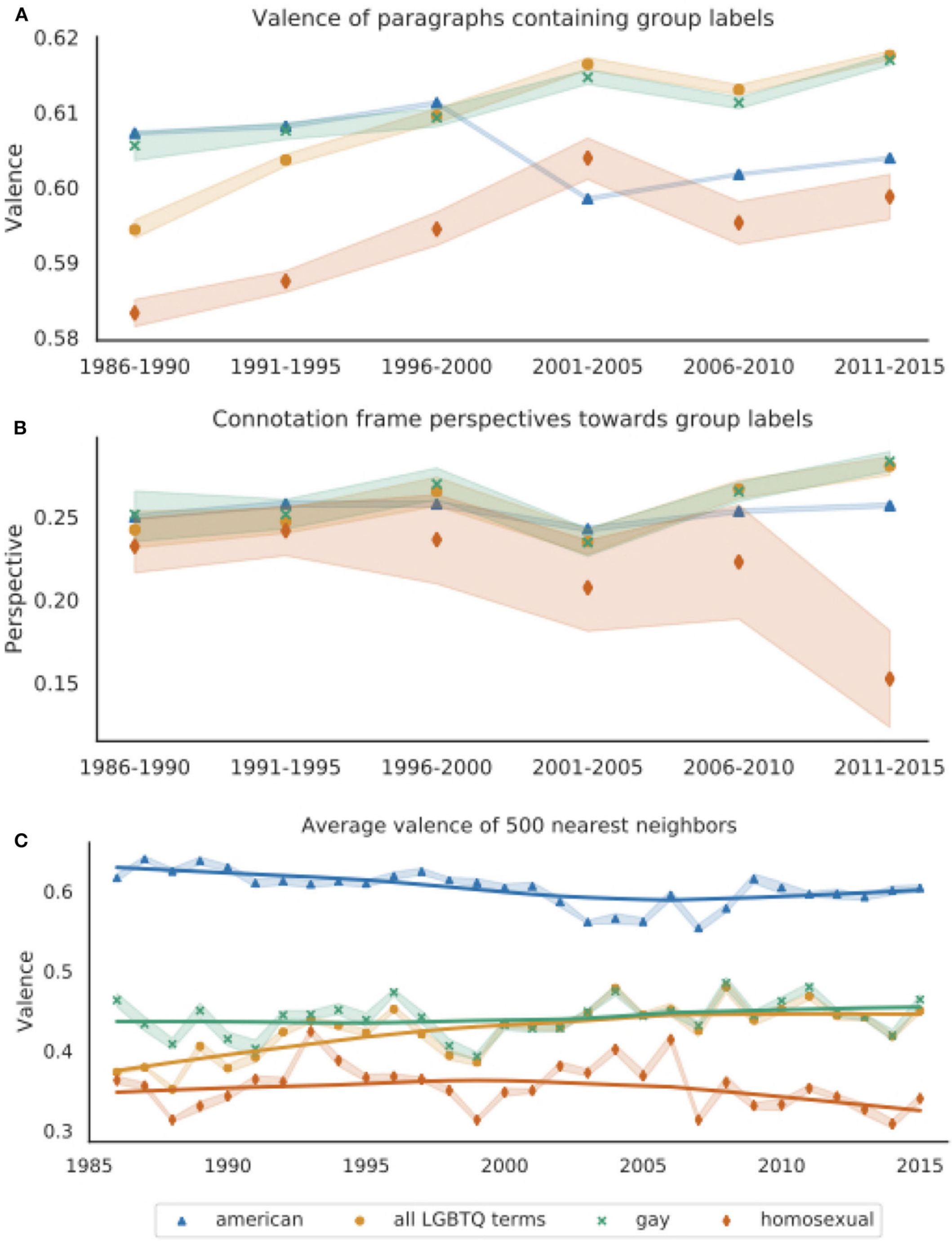


FIGURE 2 | (A) Average paragraph-level valence for paragraphs containing *gay*, *homosexual*, any LGBTQ term, and *American*, grouped into 5-years intervals. Paragraph-level scores are calculated as the average valence over all words that appear in the NRC VAD Valence Lexicon, which range from 0 (most negative) to 1 (most positive) (Mohammad, 2018). Paragraphs containing LGBTQ labels become more positive over time. Paragraphs containing *homosexual* are significantly more negative than those containing other LGBTQ labels. **(B)** Average connotation frame perspective scores over 5-years intervals. Scores are calculated for each subject-verb-object triple containing these group labels as the writer's perspective based on the head verb's entry in the Connotation Frames lexicon (Rashkin et al., 2016). **(C)** Average valence of 500 nearest words to vector representations of *gay*, *homosexual*, *all LGBTQ terms*, and *American*, averaged over 10 word2vec models trained on *New York Times* data from each year. The solid lines are Lowess curves for visualization purposes. Words' valence scores are from the NRC VAD Valence Lexicon. For all plots, the shaded bands represent 95% confidence intervals.

Connotation Frames lexicon (Rashkin et al., 2016). The wide variation, particularly for *homosexual*, is likely due to sparsity, as limiting the connotation frames analysis to verbs' immediate subject and direct object noun phrase dependents (consisting of only determiners, adjectives, and nouns) greatly reduced the amount of data for each year; there were only 39 triples for *homosexual* in 2015. We thus show results aggregated over 5-years intervals.

As with paragraph-level valence, the writer's perspective toward the label *homosexual* is significantly more negative than toward *gay* ($p < 0.001$). Linear regression indicates that perspectives toward noun phrases named by any LGBTQ term, *gay*, and *American* have all significantly increased over time ($p < 0.01$). However, the trends are still quite different, as the slopes for *gay* and *all LGBTQ terms* are an order of magnitude greater than *American* [$m = (1.1 \pm 0.39) \times 10^{-4}$ for *American*, $m = (1.4 \pm 0.18) \times 10^{-3}$ for *all LGBTQ terms*, and $m = (1.1 \pm 0.22) \times 10^{-3}$ for *gay*]. Furthermore, the writer's perspective toward noun phrases containing *homosexual* have significantly decreased over time ($p < 0.0001$).

Overall, Connotation Frames' perspective scores reveal a similar pattern as the paragraph-level valence analysis, where LGBTQ groups overall appear to be more positively evaluated in the *New York Times* over time. Unlike *gay* and the aggregated *all LGBTQ terms*, the label *homosexual* undergoes pejoration, as *homosexual* becomes increasingly used when (implicitly) expressing negative attitudes toward LGBTQ people.

5.2.1.3. Word embedding neighbor valence

Figure 2C shows the average valence scores of the 500 nearest neighbors to the vector representations of *gay*, *homosexual*, *all LGBTQ terms*, and *American* for each year. In contrast to our other techniques to quantify *negative evaluations of a target group*, this measurement notably shows that the valence of *American*'s neighboring words is significantly greater than any of the LGBTQ group representations' neighbors every year (Wilcoxon's signed-rank test, $p < 0.0001$), indicating that *American* is used in more positive contexts than LGBTQ terms. Furthermore, all LGBTQ vectors' neighbors have an average valence below the neutral 0.5. The average valence for neighboring words of *gay* and the aggregated *all LGBTQ terms* representation significantly increase over time ($p < 0.0001$), suggesting some increasing humanization in the language used in discussions of LGBTQ people.

Figure 2C also reveals dramatic connotational differences between *gay* and *homosexual*. As shown by non-overlapping confidence intervals and a Wilcoxon signed-rank test, the average valence for *homosexual*'s neighbors is significantly lower than *gay*'s neighbors over all years ($p < 0.0001$). Furthermore, while *gay*'s average neighbor valence increases over time ($p < 0.0001$), *homosexual*'s neighboring words become slightly but significantly more negative over time ($p < 0.001$). Analyzing the valence of the nearest neighbors indicates that *homosexual* has long been used in more negative (and potentially dehumanizing) contexts than *gay*, and that these words' meanings have further diverged as the label *homosexual* has been used in increasingly negative contexts over time.

5.2.2. Qualitative Analysis

5.2.2.1. Paragraph-level valence analysis

How well does paragraph-level valence analysis capture *negative evaluations of a target group*? To facilitate a qualitative evaluation of this technique, we identify several hundred paragraphs with the highest and lowest average valence. Most paragraphs with high valence scores appear to express positive evaluations of LGBTQ individuals, and those with low scores express negative evaluations.

Table 4 contain examples with extremely high and low valence. We identify several major themes from these results. Most paragraphs with high valence scores emphasize equal rights, while some focus on the activities of advocacy organizations. On the other end, paragraphs with extremely low valence often focus on violence against LGBTQ people, disease (especially AIDS), and LGBTQ issues internationally. Other themes that emerge in low-valence paragraphs include reports on (and direct quotes from) public figures who dehumanized LGBTQ people and portrayals of LGBTQ people as reckless, irresponsible, and angry.

While this technique accurately captures the valence for many paragraphs, we also identify several shortcomings. Some extreme outliers are extremely short paragraphs, including subtitles within articles which are included as paragraphs in the data. **Table 5** shows several examples that were mischaracterized by our paragraph-level valence analysis technique. In addition, there are several paragraphs with highly positive average valence that actually express negative evaluations of LGBTQ people. The valence of the third paragraph in **Table 5** is skewed by the positive words *supported* and *marriage* even though the paragraph is actually discussing low support for gay marriage. While the fourth paragraph argues that gay couples would be subpar parents relative to straight couples, it uses positive terms, such as *love* and *ideal*. Furthermore, kinship terms tend to be assigned highly positive values in the NRC VAD Valence Lexicon, including *child* and *family*. Similarly, even though the final example describes discrimination based on sexual orientation, the paragraph's average valence is impacted by positive kinship terms, such as *father* (0.812) and *mother* (0.931)[8].

Overall, our qualitative analysis shows that highly positive valence often accompanies expressions of positive evaluation toward LGBTQ groups, and low valence often accompanies expressions of negative evaluation. However, paragraph-level valence scores are also impacted by specific words cued by various topics; paragraphs about same-sex marriage tend to be more positive because words like *marriage*, *marry*, and *couple* have high valence scores while paragraphs reporting on hate crimes tend to be more negative because they contain low-valence words related to crime, violence, and injury. Furthermore, this method cannot disentangle perspectives within the text; although there are linguistic signals of dehumanization expressed in reports on anti-LGBTQ violence and homophobic speech, these dehumanizing attitudes are not necessarily from the viewpoint of the journalist or the institution. Nevertheless, there

[8] We also conducted paragraph-level sentiment analysis using binary positive vs. negative emotion lexicons, such as LIWC (Pennebaker et al., 2001), but found similar quantitative results and no qualitative improvement over the VAD lexicon.

TABLE 4 | Example paragraphs with extremely high and low valence scores, along with an interpretation of the patterns we find.

Valence	Score	Text	Year	Interpretation
High	0.853	All Americans, **gay** and non-**gay**, deserve respect and support for their families and basic freedoms.	2004	Equality
High	0.804	The experience of the joy and peace that comes with that — it was a clear indication to me that **homosexual** love was in itself a good love and could be a holy love,' Father McNeill said in the film.	2015	Equality
High	0.801	The Straight for Equality in Sports Award is given by PFLAG National, a non-profit organization for families, friends and allies of **gay**, **lesbian**, **bisexual** and **transgender** people.	2013	Advocacy
High	0.780	What do you consider the most interesting and important **LGBT** organizations working today in the city, with youth or more generally? How about more nationally?	2010	Advocacy
Low	0.266	"We kill the women. We kill the babies, we kill the blind. We kill the cripples. We kill them all. We kill the faggot. We kill the **lesbian**…When you get through killing them all, go to the goddamn graveyard and dig up the grave and kill them a-goddamn-gain because they didn't die hard enough."	1993	Direct Quote
Low	0.364	A 21-years-old college student pleaded guilty yesterday to fatally stabbing a **gay** man in Queens in what prosecutors termed a vicious burst of anti-**homosexual** violence.	1991	Violence
Low	0.403	One of his most difficult clients was a **transsexual** prostitute and drug addict who was infected with the AIDS virus and presumably spreading it to her customers and fellow addicts.	1987	AIDS
Low	0.373	Enabling promiscuity, indeed! Burroughs Wellcome is as responsible for the reckless abuse of amyl nitrate by **homosexuals** as the manufacturers of narcotic analgesics are for the horrors of opiate addiction.	1996	Recklessness
Low	0.402	The activists from Africa shrugged with resignation and sank back down on the benches. The **gay** Americans absolutely exploded at the poor woman from the airline.	2011	Recklessness
Low	0.397	Homosexuality is forbidden in Iran. Last year a United Nations report on human rights in Iran expressed concern that **gays** "face harassment, persecution, cruel punishment and even the death penalty."	2012	International

Words with extremely high valence scores (>0.85) appear in blue, and somewhat high-valence words (scores between 0.7 and 0.85) appear in light blue. Words with extremely low valence scores (<0.15) appear in red, and somewhat low-valence words (scores between 0.15 and 0.3) appear in pink. LGBTQ terms are shown in bold.

TABLE 5 | Examples mischaracterized by paragraph-level valence analysis.

Valence	Score	Text	Year	Explanation
High	0.929	Blessing of **Homosexuals**	1990	Subtitle
Low	0.031	Hate for Liberals and **Gays**	2008	Subtitle
High	0.777	Of the seven in attendance, only the Rev. Al Sharpton and Representative Dennis J. Kucinich supported **gay** marriage unambiguously.	2003	Marriage
High	0.765	And I believe children can receive love from **gay** couples, but the ideal is— and studies have shown that the ideal is where a child is raised in a married family with a man and a woman.	2005	Marriage Family
High	0.776	Ms. Bright, now a college sophomore, grew up in her mother's home but regularly visited her **gay** father, Lee, in Cartersville, Ga. She remembers when a friend was not allowed to visit her father's home because he was gay.	1993	Family

Words with extremely high valence scores (>0.85) appear in blue, and somewhat high-valence words (scores between 0.7 and 0.85) appear in light blue. Words with extremely low valence scores (<0.15) appear in red. LGBTQ terms are shown in bold.

TABLE 6 | Examples of paragraphs where the writer expresses highly positive and negative perspective toward LGBTQ groups, according to the Connotation Frames lexicon.

Perspective	Score	Text	SVO	Year
Negative	−0.83	The most forceful comment came from Cardinal Anthony J. Bevilacqua of Philadelphia, who said his archdiocese screened out gay candidates. "We feel a person who is homosexual-oriented is not a suitable candidate for the priesthood, even if **he** had never **committed any homosexual act**," the cardinal said.	S: he V: committed O: any homosexual act	2002
Positive	+0.80	"Gays are accepted here and respected here," said Mayor Tony Tarracino. "**The gays saved a lot** of the oldest parts of town, and they brought in art and culture. They deserve a lot of credit for what Key West is today."	S: the gays V: saved O: a lot	1990
Positive	+0.80	In his speech, **he praised gay rights advocates** for their hard work and also thanked many elected officials, including his predecessor, Gov. David A. Paterson, and the four Republican state senators who provided the critical votes to pass the marriage bill and whom Mr. Cuomo named one by one to some of the loudest applause of the evening.	S: he V: praised O: gay rights advocates	2011
Assigned perspective				
Negative	−0.87	Previously, Judge Vaughn Walker, who ruled the ban against same-sex unions unconstitutional in federal court, had said that ProtectMarriage could not appeal his decision to the Ninth Circuit, because they were never able to prove that **gay marriage harmed them** in any way.	S: gay marriage V: harmed O: them	2011
Positive	+0.73	Following are excerpts from opinions by the Supreme Court today in its decision that **the Constitution** does not **protect private homosexual relations** between consenting adults (…) Justice Stevens wrote a separate dissenting opinion, joined by Justices Brennan and Marshall.	S: the Constitution V: protect O: private homosexual relations	1986
Positive	+0.70	Do you know there is a Congressional candidate from Missouri who is saying that allowing **gays** into the military could **strengthen Al Qaeda**? I'm thinking, how exactly would that work? "They dance better than me, and they know how to accessorize. I'm very, very angry. It's time for jihad."	S: gays V: strengthen O: Al Qaeda	2010

Below the double line are examples of paragraphs where the writer's perspective is mischaracterized by the Connotation Frames lexicon. The relevant subject, verb, and object are shown in bold.

could be an overall dehumanizing effect if the media's discussions of a marginalized social group emphasizes such events that harm people. Repeated associations between LGBTQ labels and such negative contexts could potentially contribute to negative evaluations of LGBTQ groups.

5.2.2.2. Connotation frames of perspective

To qualitatively analyze how well the connotation frames' lexicon capture *negative evaluation of a target group*, we identify SVO tuples where the verb indicates that the writer has extremely positive or negative perspective toward either the subject or object. The first paragraph in **Table 6** contains an SVO tuple where the writer has the most negative perspective toward the noun phrases containing a group label. Inside a direct quote, this paragraph uses the phrase *any homosexual act* as the object to the verb *committed*, which has the effect of framing homosexuality as a crime. By deeming gay candidates unworthy of the priesthood, the speaker clearly negatively evaluates LGBTQ people. On the opposite end, many paragraphs labeled as containing extremely positive perspectives toward LGBTQ groups do appear to have positive evaluations of these groups. The second and third paragraphs of **Table 6** illustrate this, where *the gays* are viewed positively for having *saved* a town, and *gay rights advocates* are *praised* for their work.

However, we found several instances where paragraphs are mislabeled, shown in the bottom half of **Table 6**. In the fourth paragraph of **Table 6**, our technique identifies *gay marriage* as the subject of the negative-perspective verb *harmed*, but does not account for the preceding text, which actually contradicts the premise that *gay marriage* causes harm, and thus does not overtly negatively evaluate of LGBTQ groups (although this particular example reveals the difficulty of operationalizing this component because ProtectMarriage groups strongly oppose same-sex marriage and may have negative evaluations of LGBTQ people). The second example similarly shows that this method does not adequately account for various forms of negation, as the positive-perspective verb *protect* is actually negated. The last example in **Table 6** presents a complex case that is even challenging for qualitative analysis. Our method identifies *gays* as the subject of the verb *strengthen*, even though the subject should be the gerund *allowing gays (into the military)*, and the lexicon's entry for the writer's perspective toward the subject of *strengthen* is a highly positive 0.7. However, the object of this verb is the terrorist organization *Al Qaeda*; our background knowledge suggests that the capacity to *strengthen* Al Qaeda would reflect negative perspectives. However, this additional context provided by the rest of the paragraph indicates that the writer is being sarcastic and considers the proposition that gays have any impact on strengthening Al Qaeda to be ridiculous. Finally, the writer emphasizes their own stance in opposition to the Missouri congressional candidate by calling upon common stereotypes of gay people being good at dancing and accessorizing.

Measuring the connotation frames' lexicon perspective scores over verbs' subjects and direct objects cannot leverage as much context as measuring valence over paragraphs using the NRC VAD lexicon labeled for 20,000 words. However, this technique can make more fine-grained distinctions regarding the writer's (and institution's) attitudes directed toward LGBTQ people and is not as dramatically impacted by the emotional valence of the topic discussed. Neither technique can disentangle the journalist's perspective from those expressed by others and simply reported by the journalist. While removing direct quotations may partially address this issue, we deliberately do not remove text from direct quotes or paraphrases. The journalists and newspaper make intentional decisions about what text to include and exclude from quotations, which could still meaningfully represent their perspectives and values (Niculae et al., 2015).

5.2.2.3. Word embedding neighbor valence

Compared to the previous methods, one limitation of using word embeddings to quantify *negative evaluations of a target group* is that embeddings are not easily interpretable by analyzing a small sample of data. Instead, we assess this technique by identifying LGBTQ terms' nearest neighbors in several outlier years. To facilitate this qualitative analysis, we identify a set of *unique nearest neighbors* for each LGBTQ label in each outlier year, where a word is a unique nearest neighbor for a given LGBTQ term and year if it is not in that term's top 500 nearest neighbors in any other year.

Table 7 contains several example paragraphs that illustrate overarching themes for the outlier years 1993, 1999, and 2014. In 1999, *gay*, *homosexual* and the aggregated representation of *all LGBTQ terms* were all more closely associated with low-valence words than in almost any other year. We connect this finding to a period of intense reporting in the months following the October 1998 murder of a gay Wyoming college student, Matthew Shepard, which drew national attention to anti-LGBTQ violence. Because LGBTQ labels frequently co-occurred with text about this incident, terms related to Matthew Shepard's case had closer representations to LGBTQ terms in this year. For example, *gay* and *all LGBTQ terms*'s 500 nearest neighbors include *wyoming* in 1999 and *shepard* from the years 1998–2000. Unique nearest neighbors for *gay* in 1999 include other terms that could be connected to this incident, including *homicidal*, *imprisoned*, and *hatred*. Not only was Shepard's murder rooted in the dehumanization of LGBTQ people, but the media's emphasis on the gruesome details of Shepard's death further dehumanized him (Ott and Aoki, 2002). Ott and Aoki argue that the media's framing of this case actually further stigmatized LGBTQ people.

Our word embedding neighbor valence measure reveals that the most negative year for *gay* and *all LGBTQ terms* since 1999 was 2014, the second most-recent year of data. We identify several major themes in 2014 that co-occurred with LGBTQ group labels and possibly led to this distributional semantic pattern, primarily reporting on anti-LGBTQ laws and attitudes in Uganda and Russia (particularly in light of the 2014 Winter Olympics in Sochi). The terms *athletes* and *winterolympics* appeared in *gay*'s nearest neighbors in 2014. In addition, the terms *Uganda*, *Ugandan*, and *Mugisha* (a Ugandan LGBT advocate) are among *gay*'s unique nearest 500 neighbors in 2014.

Unlike in 1999 and 2014, LGBTQ terms in 1993 are associated with higher-valence words, especially *homosexual*. *Homosexual*'s unique nearest neighbors in 1993 include the high-valence words *pledge*, *civilian*, *readiness*, and *inclusion*. These words are likely connected with numerous stories in 1993 covering the controversy over whether LGBTQ people should be allowed to serve in the military.

5.3. Denial of Agency

5.3.1. Quantitative Results

5.3.1.1. Connotation frames of agency

Figure 3A shows the agency of each group label based on its head verb's entry in the Connotation Frames lexicon for agency (Sap et al., 2017). As in **Figure 2B**, there is large variance due to data sparsity when using the Connotation Frames lexicon, particularly for *homosexual*, which is considerably less frequent than *gay* or other LGBTQ terms in later years. In order to maximize precision when extracting subject-verb pairs, we extract only nouns and their immediate adjectival modifiers, which limits the amount of data. We thus show average agency over 5-years intervals.

Wilcoxon signed-rank tests on the means for each group labels over all years indicate that *gay* occurs in contexts with significantly higher agency than *homosexual* ($p < 0.0001$). All four group labels significantly decrease in agency over time according to linear regressions over all 30 years ($p < 0.001$), but the slope for *homosexual* is much greater [$m = (-7.9 \pm 1.3) \times 10^{-3}$ for *homosexual*, compared to $m = (-3.9 \pm .55) \times 10^{-3}$ for *gay*, and $m = (-1.5 \pm .46) \times 10^{-3}$ for *all LGBTQ terms*]. Furthermore in the most recent 15 years, *gay* and *all LGBTQ terms* show no significant change ($p = 0.097$ for *gay* and $p = 0.14$ for *all LGBTQ terms*), but *homosexual* still decreases significantly in agency ($p < 0.05$).

Figure 3A suggests that LGBTQ groups experience greater denial of agency in the *New York Times* than the institution's in-group identifier *American*. Furthermore, people described as *homosexual* experience even more denial of agency than people who are described as *gay*. Unlike the improving attitudes indicated by our analysis of *negative evaluations of a target group*, it appears that *denial of agency* increased over time for all LGBTQ groups. However, the relatively rapid decrease in agency for *homosexual* is consistent with other results suggesting *homosexual*'s pejoration.

TABLE 7 | Example paragraphs from years where LGBTQ terms' nearest neighbors had exceptionally high and low valence.

Valence	Year	Example
Low	1999	Matthew Shepard, a **gay** college student in Wyoming, had been pistol-whipped and left to die after being tied to a fence on Oct. 7, 1998. Aaron McKinney, who was charged with first-degree murder and other crimes in connection with Mr. Shepard's killing, went on trial Monday, denying that the act was a hate crime, but rather connected to drug use and outrage at a sexual advance he said Mr. Shepard made.
Low	2014	Uganda's vehement anti-**gay** movement began in 2009 after a group of American preachers went to Uganda for an anti-**gay** conference and then worked with Ugandan legislators to draft a bill that called for putting **gay** people to death. While the bill was being debated, attacks against **gay** Ugandans began to increase. In early 2011, David Kato, a slight, bespectacled man and one of the country's most outspoken **gay** rights activists, was beaten to death with a hammer.
Low	2014	"Hey, @McDonalds: You're sending #CheersToSochi while goons wearing Olympic uniforms assault **LGBT** people," read one comment last week, from the author and activist Dan Savage.
High	1993	The regulations, which are to take effect Feb. 5, codify the Administration's policy that was worked out as a compromise with the Joints Chiefs of Staff, who had defended the 50-years-old ban, arguing that allowing **homosexuals** to serve openly would hurt the morale of troops, and thus hurt military readiness.

LGBTQ terms are shown in bold.

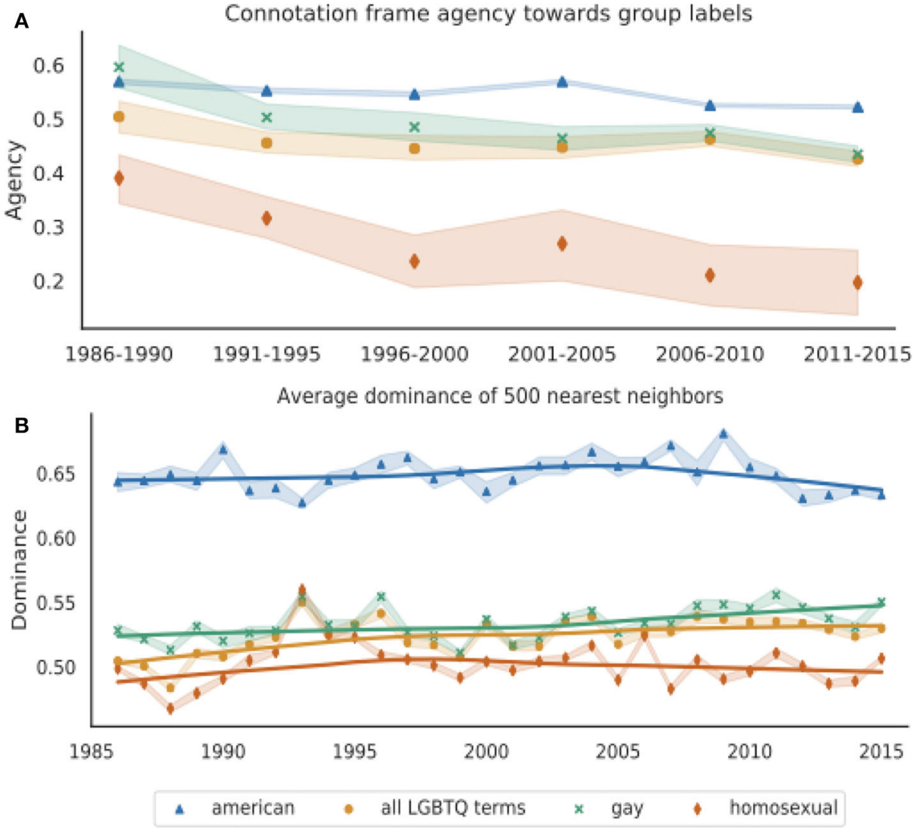


FIGURE 3 | (A) Agency of *gay*, *homosexual*, *all LGBTQ terms*, and *American* using the Connotation Frames lexicon for agency for all subject-verb-object tuples containing each group label (Sap et al., 2017), calculated over 5-years intervals. An SVO tuple received a score of 1 if the label appears in a positive agency position relative to its head verb and 0 if it does not. **(B)** Average dominance of 500 nearest words to our representations of *gay*, *homosexual*, *all LGBTQ terms*, and *American*, averaged over 10 word2vec models trained on *New York Times* data for each year. Dominance scores for each word come from the word's entry in the NRC VAD Dominance Lexicon (Mohammad, 2018), which range from 0 (least dominance) to 1 (most dominance). For both plots, the shaded bands represent 95% confidence intervals and the solid lines in **(B)** are Lowess curves for visualization purposes.

5.3.1.2. Word embedding neighbor dominance

Figure 3B shows the average dominance of each group label's 500 nearest neighbors. *American* is significantly associated with greater dominance than *gay*, *homosexual*, and *all LGBTQ terms* (Wilcoxon signed-rank test; $p < 0.0001$), and *gay* has significantly higher dominance than *homosexual* ($p < 0.0001$).

TABLE 8 | Examples where the writer attributes high and low agency toward LGBTQ groups, according to the Connotation Frames lexicon for agency.

Agency	Text	SVO	Year
High	Within the close-knit world of professional childbearers, many of whom share their joys and disillusionments online and in support groups, **gay couples** have **developed a reputation** as especially grateful clients…	S: gay couples V: developed O: a reputation	2005
High	Tonight, **the gay rights group** Stonewall Democrats will **endorse a candidate** for A.G. It's a relatively big prize in the four-man Democratic primary, given that liberal city voters will have relatively serious sway…	S: the gay rights group V: endorse O: a candidate	2006
Low	Nigeria's **gay men** and lesbians regularly **face harassment** and arrest, gay activists here say. The criminal code bans acts "against the order of nature," and imposes sentences of up to 14 years for those convicted…	S: gay men V: face O: harassment	2005
Low	Much of the debate among military and civilian officials is now focusing on some version of an approach called "don't ask, don't tell." (…) But under the "don't tell" element, there would be restrictions on the extent to which **homosexuals** could **acknowledge their homosexuality**.	S: homosexuals V: acknowledge O: their homosexuality	1993

The relevant subject, verb, and object are shown in bold.

While the dominance associated with *gay* and *all LGBTQ terms* significantly increased over time ($p < 0.0001$), the dominance associated with *homosexual* did not significantly change ($p = 0.65$). Furthermore, the average nearest neighbor dominance for *homosexual* decreased in the most recent 15 years ($p < 0.01$).

Even though dominance may more directly encode *power* rather than *agency*, the NRC VAD Dominance Lexicon is useful for operationalizing *denial of agency* because of the close relationship between these concepts. As with Connotation Frames of agency, these results suggest that LGBTQ groups experience greater *denial of agency* than the *New York Times*'s in-group *American*. Both techniques show differences between the labels *gay* and *homosexual*, where *homosexual* is consistently associated with lower agency than *gay* and further decreases over time. However, these two measurements suggest different temporal dynamics for the *denial of agency* of LGBTQ people; Connotation Frames' agency slightly decreases for *all LGBTQ terms* over time, but increases with word embedding neighbor dominance.

5.3.2. Qualitative Analysis

5.3.2.1. Connotation frames of agency

We qualitatively investigate the labels assigned by this technique for a sample of paragraphs. In general, the binary labels for positive and negative agency seem reasonably accurate, as shown by the first four example in **Table 8**. Verbs that attribute high agency to the subject include *develop* and *endorse*, suggesting that the LGBTQ-aligned subjects are in control and actively making their own decisions. On the other end, LGBTQ people have low agency when they are the subjects of passive verbs, such as *face* and *acknowledge*.

The Connotation Frames lexicon for agency seems to be especially accurate for low agency; we could not find counterexamples in our sample where LGBTQ people were portrayed with high agency but labeled with low agency. However, we found several mischaracterizations where LGBTQ people were labeled as having high agency but are not portrayed as agentive or in control of their own actions. Our Connotation Frames technique considers the example below to attribute high agency to LGBTQ people because *homosexual* appears in the subject of the high-agency verb *violate*; however, *homosexual* actually modifies *relationships*, not people themselves. Furthermore, this debate within religion appears to be devoid of input from LGBTQ people and does not portray them as particularly agentive.

- At the same time, it underscored a stark division in Judaism over the place of homosexuals in society. Orthodox rabbinical groups believe that **homosexual relationships violate Jewish law**…(1996)

5.3.2.2. Word embedding neighbor dominance

Using the VAD Dominance Lexicon to calculate average dominance of each social group label corresponds well to our notion of *denial of agency*. Because *gay*'s nearest neighbors have a much higher average dominance than *homosexual*'s for most years, we compare words that are nearby neighbors for *gay* and not *homosexual* for multiple years' word2vec spaces. Words frequently among the 500 words nearest to *gay* and not *homosexual* include high-agency words, such as *activist*, *liberation*, *advocate*, and *advocacy*, which have dominance scores of 0.877, 0.857, 0.818, and 0.731, respectively. Words frequently among *homosexual*'s 500 nearest neighbors and not *gay*'s include low-agency words, such as *submissive* (0.173), *degrading* (0.232), *enslavement* (0.302), and *repressed* (0.311).

We additionally investigate the word2vec models corresponding to several outlier years. *Homosexual*'s neighbors have the highest average dominance in 1993, which is likely due to military-related language in debates surrounding the "Don't Ask, Don't Tell" legislation. High-dominance words unique to *homosexual*'s nearest neighbors in 1993 include *forces* (0.886), *military* (0.875), *enforce* (0.836) and *troops* (0.804). *Gay*'s neighbors' in 1999 have the lowest average dominance than any other year, which is likely connected to Matthew Shepard's death and the subsequent outrage; unique neighbors for *gay* in 1999 include *imprisoned* (0.302) and *repressed* (0.311).

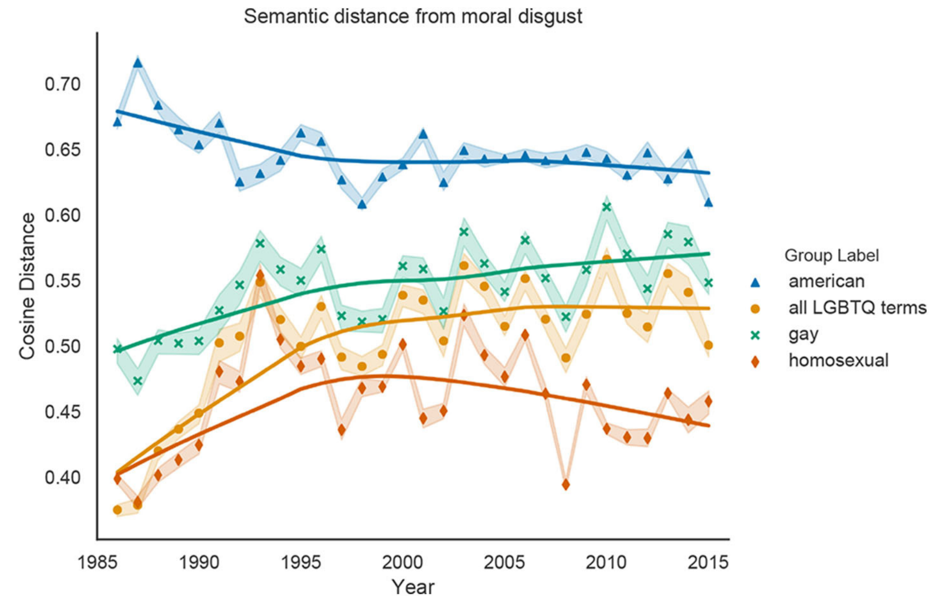


FIGURE 4 | Cosine distance between our representations of *gay*, *homosexual*, *all LGBTQ terms*, and *American* and the vector representation of the *Moral Disgust* concept, averaged over 10 word2vec models trained on *New York Times* data for each year. Increases in cosine distance indicate decreases in *Moral Disgust*; possible values range from 0 (most closely associated with Moral Disgust) to 1 (least associated with Moral Disgust). Shaded bands represent 95% confidence intervals and the solid lines are Lowess curves for visualization purposes.

5.4. Moral Disgust

5.4.1. Quantitative Results

Figure 4 shows the changing relationships between *all LGBTQ terms*, *gay*, *homosexual* and the dehumanizing concept of *Moral Disgust*. Because the cosine distance between *American* and *Moral Disgust* is significantly greater over all years than any LGBTQ representation (Wilcoxon signed-rank test; $p < 0.0001$), *American* is the least associated with *Moral Disgust*. Furthermore, the cosine distance between *gay* and *Moral Disgust* is significantly greater than the distance between *homosexual* and *Moral Disgust* for every year ($p < 0.0001$), indicating that *homosexual* is more closely associated with *Moral Disgust* than *gay* is. Linear regression analyses show that *all LGBTQ terms* and *gay* significantly increase in cosine distance from the *Moral Disgust* vector ($p < 0.0001$), indicated weakening associations between LGBTQ people and moral disgust over time. On the other hand, the distance between *homosexual* and *Moral Disgust* does not change significantly over time ($p = 0.54$), and even decreases after 2000 ($p < 0.05$).

Overall, these measurements of associations between LGBTQ people and *Moral Disgust* are consistent with our other operationalizations of dehumanization. All LGBTQ labels are more closely associated with *Moral Disgust* than the newspaper's in-group term *American*, but these associations weaken over time, suggesting increased humanization. Notably, the term *homosexual* has always been more associated with *Moral Disgust* than the denotationally-similar term *gay*, and *homosexual* actually becomes more closely associated with this dehumanizing concept in recent years.

5.4.2. Qualitative Analysis

Our analysis of *homosexual*'s changing semantic neighbors from **Table 3** has shown that this term has become more associated with immoral concepts, suggesting that moral disgust is a mechanism by which LGBTQ people are dehumanized. Although rarely directly invoked, the connection between LGBTQ people and disgust is supported by the data, such as in the examples shown below, where words belonging to the moral disgust lexicon are in bold. **Figure 4** indicates that late 1980s and early 1990s, LGBTQ labels rapidly became more semantically distant from *Moral Disgust*. This likely reflects decreasing attention to the AIDS epidemic, as many disease-related words are included in the moral disgust lexicon.

- Senator Jesse Helms, the North Carolina Republican who has vigorously fought homosexual rights, wants to reduce the amount of Federal money spent on AIDS sufferers, because, he says, it is their "deliberate, **disgusting**, revolting conduct" that is responsible for their **disease** (1995).
- A lawyer named G. Sharp, address unknown, called the cover picture "utterly **repulsive**." Donald Ingoglia of Sacramento was equally outraged. "Showing two smiling gays on the cover illustrates how **sick** our society has become," he wrote. "You have my non-lawyer friends falling off their chairs" (1992).
- …Mr. Robison could be harsh—he yelled in the pulpit and referred to gay men and lesbians as **perverts**—but Mr. Huckabee was a genial ambassador … (2008)
- …When bishops started telling parishioners that their gay and lesbian siblings were **sinners**, and that family planning was a

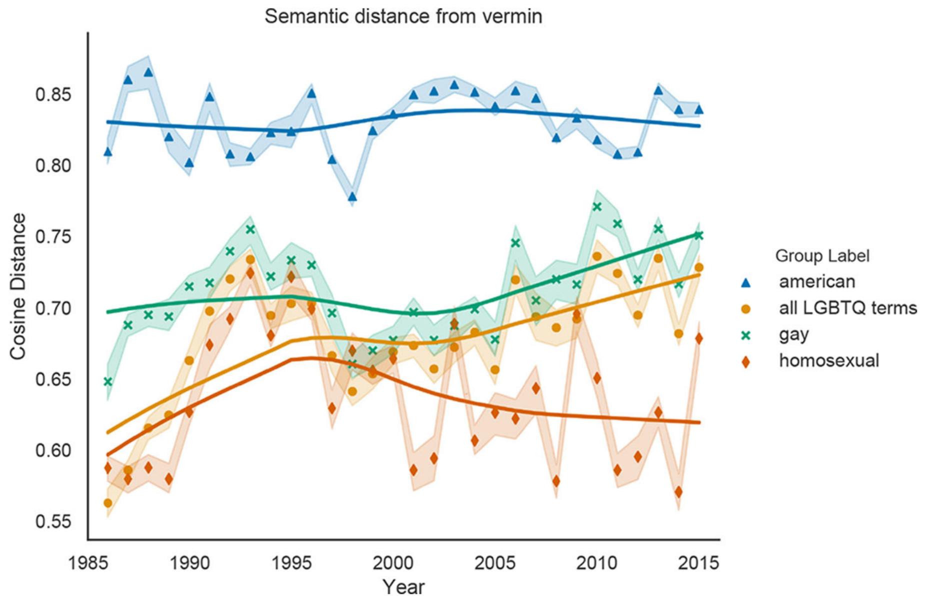


FIGURE 5 | Cosine distance between our representations of *gay*, *homosexual*, *all LGBTQ terms*, and *American* and the vector representation of the *Vermin* concept, averaged over 10 word2vec models trained on *New York Times* data for each year. Possible values for cosine distance range from 0 (most closely associated with *Vermin*) to 1 (least associated with *Vermin*). Shaded bands represent 95% confidence intervals, and the solid lines are Lowess curves for visualization purposes.

grievous wrong, people stopped listening to them—for good reason (2013).

5.5. Vermin as a Dehumanizing Metaphor

5.5.1. Quantitative Results

Figure 5 shows the relationships between LGBTQ labels (and *American*) and the dehumanizing vermin metaphor, quantified as the cosine distance between the labels' word2vec vectors and a *Vermin* concept representation, which is the centroid of multiple vermin-related words. As with *Moral Disgust*, the in-group term *American* is further away from *Vermin* over all years than any LGBTQ term (Wilcoxon signed-rank test; $p < 0.0001$). The cosine distance between *gay* and *Vermin* is also greater than between *homosexual* and *Vermin* ($p < 0.0001$), indicating that *homosexual* is more closely associated with the dehumanizing vermin metaphor than *gay* is. Furthermore, while *all LGBTQ terms* and *gay* become more semantically distant from *Vermin* over time, ($p < 0.0001$), the association between *Vermin* and *homosexual* does not significantly change over time ($p = 0.13$).

This measure of the implicit *vermin metaphor* reveals similar patterns as the other dehumanization measures. Overall, LGBTQ groups are more associated with vermin than the comparison group *American*, but this association weakens over time, suggesting increased humanization. In addition, *homosexual* has become a more dehumanizing term, with stronger associations with vermin than other LGBTQ labels.

5.5.2. Qualitative Analysis

Metaphors comparing humans to vermin have been especially prominent in dehumanizing groups throughout history (Haslam, 2006; Steuter and Wills, 2010). Although no New York Times writers directly compare LGBTQ people to vermin, this metaphor may be invoked in more subtle ways. There are only three paragraphs in the *LGBTQ corpus* that explicitly mention vermin in order to criticize the LGBTQ people-as-vermin metaphor. Nevertheless, these paragraphs point to the existence of this metaphor.

- Since gay women can't be stigmatized en masse with AIDS, the council had to use real ingenuity to prove that they, too, are vermin at "much greater risk from one another" than from gay-bashers . . . (1998)
- "The equating of gay men to vermin is appalling," Addessa said from Philadelphia. "We need to encourage the Eagles and Owens to make a public apology and for the Eagles to publicly discipline Owens. These comments that equate gay men to some inferior life form do real harm, creating a cultural environment which justifies violence against gay and lesbian people (2004).
- In 3 h at training camp Tuesday, he hustled vigorously through practice, eagerly signed autographs for visiting military personnel and tried to explain incendiary remarks that appeared in a magazine regarding the sexual orientation of a former teammate in San Francisco, words that seemed to compare gays to rodents (2004).

6. HUMAN EVALUATION OF VECTOR-BASED MEASURES

Our vector-based methods can directly capture associations between LGBTQ people and dehumanizing concepts. However, findings from these methods are difficult to interpret, as discussed in earlier qualitative analysis sections. Furthermore, while the NRC VAD Lexicon and the Connotation Frames Lexicons have been evaluated in prior work (Rashkin et al., 2016; Sap et al., 2017; Mohammad, 2018), our vector-based methods have not. Thus, we recruit humans from Amazon Mechanical Turk (MTurk) to quantitatively evaluate our four vector-based measures: word embedding neighbor valence (for *negative evaluation of a target group*), word embedding neighbor dominance (for *denial of agency*), semantic distance from the concept of *moral disgust*, and semantic distance from the concept of *vermin*.

Although these four measures rely on vector representations of LGBTQ labels and not individual paragraphs, we use paragraphs as the unit of analysis for our evaluation in order for the task to be feasible for human annotators. In section 6.1, we describe how we use our vector-based methods to obtain the most and least dehumanizing paragraphs for each dehumanization component. We discuss the MTurk task design in section 6.2 and results in section 6.3.

6.1. Identifying the Most (De)humanizing Paragraphs

6.1.1. Word Embedding Neighbor Valence and Dominance

Our word embedding neighbor valence and dominance methods are proxies for measuring the *negative evaluation of the target group* and *denial of agency* dimensions of dehumanization, respectively. They directly estimate the valence and dominance scores for LGBTQ terms based on NRC VAD entries for each term's semantic neighbors.

To obtain full paragraphs corresponding to the most and least dehumanizing extremes of *negative evaluation of a target group*, we first train word2vec on the entire *New York Times* dataset using the same hyperparameters as in section 3.1.3. Let N be the nearest 500 words to the representation of *all LGBTQ terms* in this vector space, and let V and D be the full NRC Valence and Dominance Lexicons. We define subset lexicons, $V_s = N \cap V$ and $D_s = N \cap D$; V_s and D_s are the subsets of the NRC Valence and Dominance Lexicons containing only words that neighbor *all LGBTQ terms*. We calculate *neighbor valence* scores for each paragraph P as $\frac{1}{|P|}\Sigma_{w \in P} V_s[w]$, where $|P|$ is the total number of tokens in P and $V_s[w]$ is the valence score of w. Similarly, we calculate *neighbor dominance* scores as $\frac{1}{|P|}\Sigma_{w \in P} D_s[w]$.

For human evaluation, we consider paragraphs with the highest and lowest scores for *neighbor valence* and *neighbor dominance*. We remove paragraphs containing fewer than 15 or more than 75 words. Because our case study focuses on the words *gay(s)* and *homosexual(s)*, we further restrict our sample to paragraphs containing these terms.

6.1.2. Moral Disgust and Vermin Metaphor

We measure implicit associations of LGBTQ groups with *moral disgust* and *vermin* by calculating the cosine distance between LGBTQ terms' vectors and vector representations of *moral disgust* and *vermin*. Thus, we identify paragraphs corresponding to the most and least dehumanizing extremes by comparing the cosine distance between paragraph embeddings and the *Moral Disgust* and *Vermin* concept vectors. We create each paragraph's embedding by calculating the tfidf-weighted average of all words' vectors and removing the first principal component, which improves the quality of sentence and document embeddings (Arora et al., 2019).

We select the paragraphs that are the closest (most semantically similar) and furthest from the *Moral Disgust* and *Vermin* vectors based on cosine distance. As in section 6.1.1, we limit our sample to paragraphs containing between 15 and 75 words and either the term *gay(s)* or *homosexual(s)*.

6.2. MTurk Task Design

As discussed in our qualitative analyses, journalistic text captures numerous perspectives, not only from journalists themselves, but also from people quoted and people or groups described within the text. While our current computational methods do not disambiguate these perspectives, human evaluation can provide insights into whose perspectives primarily drive our findings about dehumanization. Thus, we manually divide each measure's most and least dehumanizing paragraphs into three categories based on whose views are most prominent: the author, a person quoted or paraphrased, or a person/group mentioned or described within the text. For each measure, our final sample for human evaluation consists of the 20 most humanizing and 20 most dehumanizing paragraphs within each of the three "viewpoint" categories, yielding 120 paragraphs for each vector-based measure.

MTurk workers read a paragraph and answered a question about the attitudes of the author, person quoted, or people mentioned/described in the text. **Table 9** shows four examples, the dehumanization component that they correspond to, whether they are ranked high (most dehumanizing) or low (least dehumanizing), the most prominent viewpoint, and the exact question that workers answered. The question depends on which dehumanization component's measure is being evaluated. In addition, we include the actual name of people quoted or mentioned in order to simplify the task. Each question is answered with a 5-point Likert scale with endpoints specified in the task. For the *negative evaluation* and *denial of agency* questions, 1 is the most dehumanizing option and 5 is the most humanizing option, but the opposite is the case for *vermin* and *moral disgust*. As a post-processing step, we reverse the scale for the latter so higher values always correspond to more humanizing views.

Three MTurk workers completed each task. Workers were located in the United States, already completed at least 1,000 MTurk tasks, and have an approval rate of at least 98%. Each task took ∼20–25 s and workers were compensated $0.05. To avoid confusion with multiple question formulations, we published the tasks for each dehumanization component separately.

6.3. Human Evaluation Results

The results from the MTurk study, shown in **Figure 6**, largely support our use of vector-based measures. Paragraphs with the highest *neighbor valence* were judged to hold more positive evaluations of gay people ($p < 0.0001$). Paragraphs whose

TABLE 9 | Examples of four paragraphs annotated by MTurk workers, one for each dehumanization component.

Paragraph	Component	Extreme	Viewpoint	Question
Some people think that equality can be achieved by offering gays civil unions in lieu of marriage. Civil unions are not a substitute for marriage. Separate rights are never equal rights.	Negative evaluation	Low	Author	How does the author feel about gay people?
"I also learned it was possible to be black and gay," Mr. Freeman said. "The first black gay I met, I didn't believe it. I thought you could only be a member of one oppressed minority."	Denial of agency	High	Person quoted	To what extent does Mr. Freeman think that gay people are able to control their own actions and decisions?
In a speech exceptional for its deep emotion and sharp message, Ms. Fisher implicitly rebuked those in her party who have regarded the sickness as a self-inflicted plague earned by immoral behavior—homosexual sex or intravenous drug abuse.	Moral disgust	High	Person mentioned	To what extent does Ms. Fisher's party consider gay people to be disgusting or repulsive?
The Supreme Court on Tuesday was deeply divided over one of the great civil rights issues of the age, same-sex marriage. But Justice Anthony M. Kennedy, whose vote is probably crucial, gave gay rights advocates reasons for optimism based on the tone and substance of his questions.	Vermin	Low	Person mentioned	Vermin are animals that carry disease or cause other problems for humans. Examples include rats and cockroaches. To what extent does [the author] consider gay people to be vermin-like?

Extreme refers to whether the paragraph is ranked as the most dehumanizing (high) or least dehumanizing (low) for each measure. Viewpoint refers to whose perspective workers are asked to reason about. The question that MTurk workers answer is modified based on both the dehumanization component and the viewpoint.

embeddings are nearest to the *Moral Disgust* concept vector are judged to express stronger views of gay people as "disgusting" or "repulsive" compared to the furthest paragraphs ($p < 0.0001$). Similarly, paragraphs nearest to *Vermin* concept consider gay people to be more vermin-like than the paragraphs furthest away ($p < 0.0001$)[9].

The only component that does not follow these expected results is *denial of agency*, where paragraphs with highest and lowest *neighbor dominance* are not judged to be significantly different ($p = 0.19$). This may reflect that using a lexicon for dominance is not a perfect proxy for the more nuanced concept of agency. Another possible explanation is the inherent complexity in measuring *denial of agency*. While the other components are already challenging by requiring an annotator to reason about another person's attitudes toward the target group, assessing *denial of agency* is even more complicated, as it requires an annotator to reason about another person's perceptions of the cognitive capabilities of members of the target group.

The bottom row of **Figure 6** separates the results based on whose viewpoint MTurk workers are asked to reason about: the paragraph's author, the people quoted, or the people mentioned in the text. This reveals a strikingly consistent pattern; the difference between the two extremes is largest when workers are asked about the *people mentioned*, smallest when asked about *the author*, and in-between when asked about *people quoted*. This suggests that dehumanizing representations of LGBTQ people in the *New York Times* may be most driven by descriptions about other people's attitudes, and to a lesser extent, direct quotes and paraphrases.

7. DISCUSSION

Our framework for the computational linguistic analysis of dehumanization involves identifying major dimensions of dehumanization from social psychology literature, proposing linguistic correlates for each dimension, and developing robust and interpretable computational methods to quantify these correlates. We apply this framework to study the dehumanization of LGBTQ people in the *New York Times* from 1986 to 2015. We measure four dimensions of dehumanization: *negative evaluations of a target group*, *denial of agency*, *moral disgust*, and (implicit) invocations of *vermin metaphors*. Overall, we find increasingly humanizing descriptions of LGBTQ people over time. LGBTQ people have become more associated with positive emotional language, suggesting that *negative evaluations of the target group* have diminished. LGBTQ people have become more associated with higher-dominance words, suggesting decreasing *denial of agency*, although this finding was not replicated with the verb-centric "Connotation Frames" measurement. Furthermore, labels for LGBTQ people have moved further away from the concepts of *moral disgust* and *vermin* within distributional semantic representations, suggesting that harmful associations between LGBTQ people and these dehumanizing concepts have weakened over time.

Despite these trends, the labels *gay* and *homosexual* exhibit strikingly different patterns. *Homosexual* is associated with more negative language than *gay*, suggesting more negative evaluations of people described as *homosexual* than *gay*.

[9] We evaluate our methods in this way instead of using traditional precision and recall metrics because annotators rated each example on a 5-point scale, so binarizing annotations risks losing valuable information. Precision, recall, and F1 scores for each component can be found in the **Supplementary Material**.

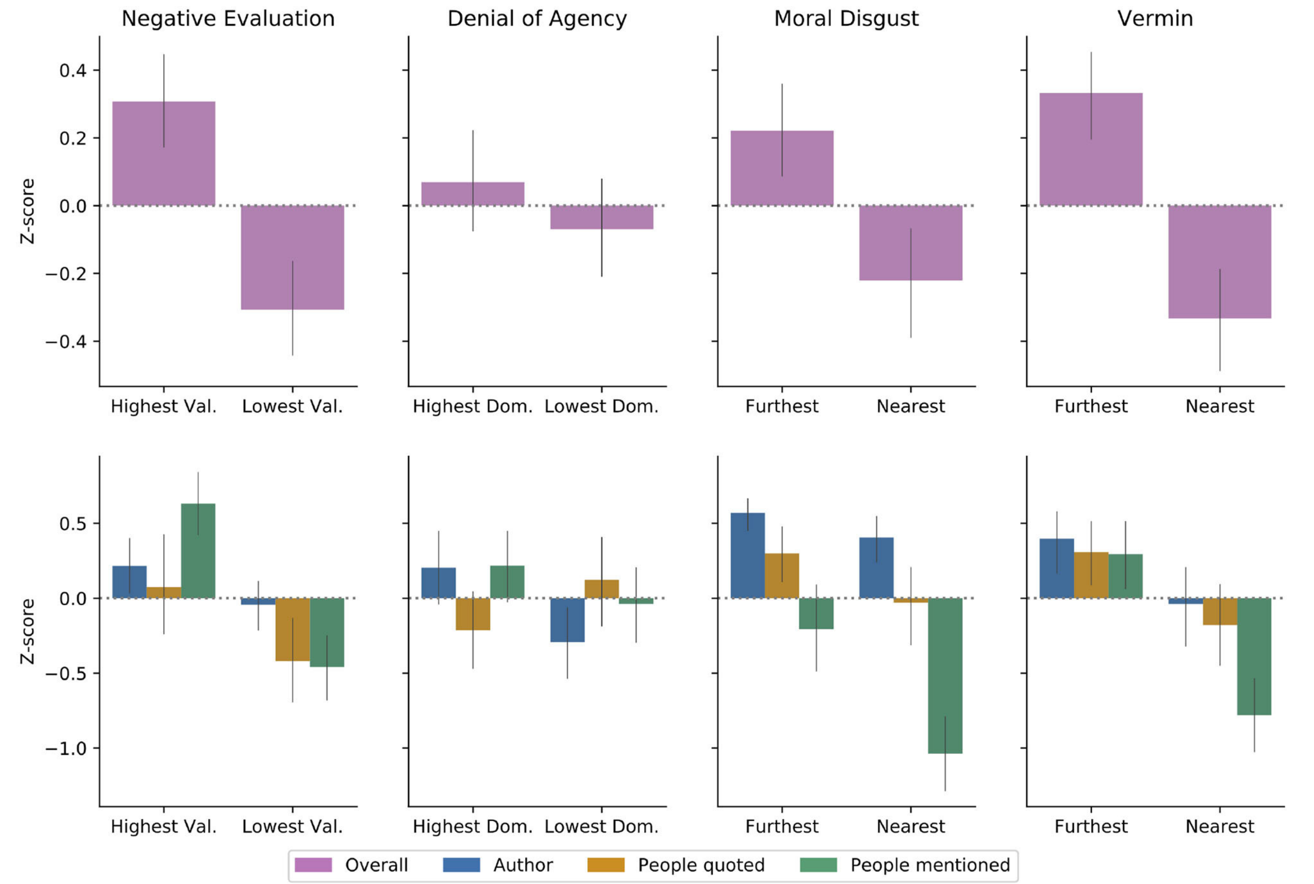


FIGURE 6 | Results from human evaluation of our vector-based methods for quantifying *negative evaluation of the target group*, *denial of agency*, *moral disgust*, and *vermin metaphor*. Higher values are more humanizing (more positive evaluation, greater agency, less association with moral disgust or vermin) and lower values are more dehumanizing. The top row shows overall ratings after z-score normalization for each component and the bottom row separates ratings by the viewpoint workers are asked to judge.

Homosexual is also associated with greater *denial of agency*, and has stronger connections to *moral disgust* and *vermin* than *gay*. Unlike for other LGBTQ labels, discussions of *homosexual* people have not become more humanizing over time, and several techniques even suggest that *homosexual* has become used in more dehumanizing contexts since 2000. Through its repeated use in these contexts, the use of the word *homosexual* appears to have emerged as an index of more dehumanizing attitudes toward LGBTQ people than other labels. Despite the denotational similarity between *homosexual* and *gay*, our computational techniques tracks the stark divergence in social meanings.

We restrict our analysis to the lexical level for ease of interpretability, and leveraged a diverse array of existing resources, including the NRC VAD lexicon (Mohammad, 2018), Connotation Frames lexicons (Rashkin et al., 2016; Sap et al., 2017), and the Moral Foundations Dictionary (Graham et al., 2009). For *negative evaluations of a target group* and *denial of agency*, we propose multiple different techniques that vary in accuracy and interpretability. Word-counting methods are often inaccurate due to their simplicity but their results are easily interpretable, while embedding-based methods suffer the opposite problem. Carefully considering the tradeoff between model quality and interpretability is especially important in work that seeks to characterize complex and sensitive social phenomena, such as dehumanization.

7.1. Limitations and Future Work

As the first attempt to computationally analyze dehumanization, this work has many limitations. While we demonstrate how the proposed techniques capture linguistic signals of dehumanization, our qualitative and quantitative evaluation suggest that the findings may be driven more by events and attitudes of people described in the text rather than the journalists' own views. An exciting area of future work could involve developing more sophisticated methods to disambiguate the writer's attitudes, attitudes of people mentioned or quoted, and events, while recognizing that each of these could contribute to the overall representation of marginalized groups in the media. In addition, the present work uses word2vec since all known affective lexicons are type-level, but contextualized embedding-based methods have great potential for more nuanced analyses of dehumanizing language by leveraging token-level representations (Devlin et al., 2018; Peters et al., 2018).

Our framework could be expanded to include more insights from dehumanization theory. Beyond the four components discussed in this article, social psychology research has identified other cognitive processes that contribute to dehumanization, including *psychological distancing*, *essentialism* (the perception that the target group has some essence that makes them categorically and fundamentally different), and *denial of subjectivity* (neglect of a target group member's personal feelings and experiences) (Rothbart and Taylor, 1992; Nussbaum, 1999; Graf et al., 2013; Haslam and Stratemeyer, 2016). Scholars also differentiate between two forms of dehumanization, *animalistic* (likening humans to animals) and *mechanistic* (likening humans to inanimate objects or machines), which may differ substantially in their linguistic expressions (Haslam, 2006).

For simplicity and ease of interpretation, we quantify lexical cues of dehumanization. However, our understanding of dehumanizing language would be enriched by considering linguistic features beyond the lexicon. For example, Acton (2014) has shown that definite plurals in English (e.g., *the gays*) have a unique social and pragmatic effect compared to bare plurals (e.g., *gays*) by packaging individual entities into one monolith and setting this group apart from the speaker. Indexing a speaker's non-membership in the group being discussed creates social distance between the speaker and group (Acton, 2014), which makes it likely that using definite plurals to label marginalized social groups plays an important role in dehumanization. Similarly, examining non-lexical signals could help us capture elements of dehumanization not easily identifiable with lexical resources alone. For example, a group label's word class (e.g., *gay* as a noun or adjective) may have implications for *essentialism*, as adjectives simply name attributes to some entity, while nouns explicitly state the entity's category membership and encapsulates other stereotypes associated with that category (Wierzbicka, 1986; Hall and Moore, 1997; Graf et al., 2013; Palmer et al., 2017). We furthermore believe that incorporating discourse-level analysis, such as examining the role of direct quotes in an article and who is being quoted, could provide informative insights that could address some limitations discussed earlier.

We support our proposed framework with a case study of LGBTQ representation in the *New York Times*. This case study is limited as an analysis of the dehumanization of LGBTQ people in the media. We only investigate one data source, which does not capture the entirety of media discourse about LGBTQ people. Furthermore, we only study newspaper articles written in (Standard) American English. Future work could focus on cross-linguistic comparisons of dehumanizing language and assess how well our measures generalize to other languages. Finally, the case study focuses on the labels *gay* and *homosexual* due to data availability. As a consequence, we have less understanding about the differences and changes in representations of LGBTQ people who do not identify with these labels.

The primary aim of this paper is to develop a computational framework for analyzing dehumanizing language toward targeted groups. While our in-depth case study focuses on one particular social group, this framework can be generalized to study dehumanization across a wide variety of social groups, and this could be a fruitful area of future work. For example, Asians have faced increased prejudice and dehumanization since the beginning of the COVID-19 pandemic (Van Bavel et al., 2020; Vidgen et al., 2020; Ziems et al., 2020). Our framework could be applied to understand who dehumanizes these populations in both news and social media, and how the degree of dehumanization changes over time or varies by region. This framework could provide a nuanced view into the shifting nature of dehumanization toward Asians. For example, the "Asians are good at math" stereotype may have led dehumanization via *denial of agency* or *denial of subjectivity* (Shah, 2019). However, stereotypes of Asians as COVID-19 carriers may have made *moral disgust* and *associations with vermin* more salient mechanisms of dehumanization. In our case study, we use computational measures of dehumanizing language to show how the terms *gay* and *homosexual* have diverged in meaning. This method of demonstrating how denotationally similar items differ in connotation can also generalize to other issues and social groups. For example, we may expect labeling COVID-19 as the *Wuhan Virus* or *Chinese Virus* may be associated with greater dehumanization of Asians than the names *COVID-19* or *SARS-CoV-2* (Van Bavel et al., 2020; Xu and Liu, 2020).

7.2. Ethical Implications

We hope to draw attention to issues of media representation of marginalized groups and to eventually help make the online world a safer and kinder place. An important part of this mission is to acknowledge the ethical implications and potential issues of our own work.

The methods that we use to quantify dehumanization are themselves biased and potentially harmful. For example, we show in section 5.2.1.1 that the lexicon used to measure valence contains its own anti-LGBTQ biases by considering LGBTQ group labels to be primarily negative/unpleasant. We also train word2vec models on *New York Times* data, which presents biases. Though models trained on biased data are typically concerning due to harmful downstream effects (Bolukbasi et al., 2016), we leverage this data as a resource for uncovering human biases and understanding *how* biases emerge in the media.

Another concern of this work in our computational methods to represent human beings. Representing people as sequences of numbers (especially in our vector-based experiments) is inherently dehumanizing. While we hope that this work will humanize and empower marginalized groups, we acknowledge that it can also have effect of perpetuating their dehumanization.

Other ethical implications of this project appear within our case study. We do not include LGBTQ labels, such as *queer* or *trans*, which often had different meanings and were found in unrelated contexts in earlier years. Furthermore, our analysis uses an aggregated representation for LGBTQ people, which unintentionally minimizes the vast diversity of social identities encompassed within this umbrella. We highlight striking temporal changes and differences between *gay* and *homosexual*, which were chosen because these labels were well-represented in all years. However, emphasizing these labels at the expense of others may contribute to the erasure of people who are marginalized even within LGBTQ communities.

8. CONCLUSION

This work is the first known computational linguistic study of dehumanization, and provides contributions to multiple fields. The proposed framework and techniques to quantify salient components of dehumanization can shed light on linguistic variation and change in discourses surrounding marginalized groups. Furthermore, these tools for large-scale analysis have potential to complement smaller-scale psychological studies of dehumanization. Finally, this work has implications for automatically detecting and understanding media bias and abusive language online.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JM, YT, and DJ collaborated on the conception and design of the study, read, and revised the manuscript. JM prepared the data, conducted the case study analysis, and conducted statistical analysis. A first draft of the paper was written by JM. All authors approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank Rob Voigt, Chris Potts, Penny Eckert, David Jurgens, Ceren Budak, Anna-Marie Sprenger, and Ismael Mendoza for their insightful contributions and feedback. We would also like to thank our audience at NWAV-46, who provided helpful feedback on an earlier version of this project. Finally, we thank our reviewers for their thoughtful and valuable comments. This research was supported in part by Public Interest Technology University Network Grant No. NVF-PITU-Carnegie Mellon University-Subgrant-009246-2019-10-01.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00055/full#supplementary-material

REFERENCES

Acton, E. K. (2014). *Pragmatics and the social meaning of determiners* (Ph.D. thesis), Stanford University, Stanford, CA, United States.

Alorainy, W., Burnap, P., Liu, H., and Williams, M. L. (2019). "The enemy among us" detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web* 13, 1–26. doi: 10.1145/3324997

Arora, S., Liang, Y., and Ma, T. (2019). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017.* (Toulon).

Barnhurst, K. G., and Mutz, D. (1997). American journalism and the decline in event-centered reporting. *J. Commun*. 47, 27–53. doi: 10.1111/j.1460-2466.1997.tb02724.x

Bar-Tal, D. (1990). Causes and consequences of delegitimization: models of conflict and ethnocentrism. *J. Soc. Issues* 46, 65–81. doi: 10.1111/j.1540-4560.1990.tb00272.x

Baumer, E., Elovic, E., Qin, Y., Polletta, F., and Gay, G. (2015). "Testing and comparing computational approaches for identifying the language of framing in political news," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO), 1472–1482. doi: 10.3115/v1/N15-1171

Blodgett, S. L., and O'Connor, B. (2017). Racial disparity in natural language processing: a case study of social media african-american english. *arXiv[Preprint].arXiv*:1707.00061

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona), 4356–64.

Boydstun, A. E., Card, D., Gross, J., Resnick, P., and Smith, N. A. (2014). *Tracking the Development of Media Frames Within and Across Policy Issues*.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). "Identifying media frames and frame dynamics within and across policy issues," in *New Directions in Analyzing Text as Data Workshop* (London).

Breitfeller, L., Ahn, E., Jurgens, D., and Tsvetkov, Y. (2019). "Finding microaggressions in the wild: a case for locating elusive phenomena in social media posts," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong: Association for Computational Linguistics), 1664–1674. doi: 10.18653/v1/D19-1176

Buckels, E. E., and Trapnell, P. D. (2013). Disgust facilitates outgroup dehumanization. *Group Process. Intergr. Relat*. 16, 771–780. doi: 10.1177/1368430212471738

Burnap, P., and Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Sci*. 5:11. doi: 10.1140/epjds/s13688-016-0072-6

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Card, D., Boydstun, A., Gross, J. H., Resnik, P., and Smith, N. A. (2015). "The media frames corpus: annotations of frames across issues," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Beijing), 438–444. doi: 10.3115/v1/P15-2072

Demszky, D., Garg, N., Voigt, R., Zou, J., Shapiro, J., Gentzkow, M., et al. (2019). "Analyzing polarization in social media: method and application to tweets on 21 mass shootings," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN), 2970–3005. doi: 10.18653/v1/N19-1304

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv[Preprint].arXiv:* 1810.04805.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst*. 2:30. doi: 10.1145/2362394.2362400

Dinu, G., Lazaridou, A., and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv[Preprint].arXiv:* 1412.6568.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018). "Hate lingo: a target-based linguistic analysis of hate speech in social media," in *Twelfth International AAAI Conference on Web and Social Media* (Palo Alto, CA).

Entman, R. M. (1993). Framing: toward clarification of a fractured paradigm. *J. Commun*. 43, 51–58. doi: 10.1111/j.1460-2466.1993.tb01304.x

Esses, V. M., Medianu, S., and Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *J. Soc. Issues* 69, 518–536. doi: 10.1111/josi.12027

Fast, E., and Horvitz, E. (2016). Long-term trends in the public perception of artificial intelligence. *arXiv[Preprint].arXiv:*1609.04904.

Field, A., Bhat, G., and Tsvetkov, Y. (2019). "Contextual affective analysis: a case study of people portrayals in online# metoo stories," in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13 (Munich), 158–169.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). "Framing and agenda-setting in russian news: a computational analysis of intricate political strategies," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 3570–3580. doi: 10.18653/v1/D18-1393

Field, A., and Tsvetkov, Y. (2019). "Entity-centric contextual affective analysis," in *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (Florence). doi: 10.18653/v1/P19-1243

Gallup (2019). *Gay and Lesbian Rights*. Available online at: http://news.gallup.com/poll/1651/gay-lesbian-rights.aspx

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A*. 115, E3635–E3644. doi: 10.1073/pnas.1720347115

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., and Dehghani, M. (2016). "Morality between the lines: detecting moral sentiment in text," in *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes* (New York, NY). Retrieved from: http://mortezadehghani.net/wp-content/uploads/morality-lines-detecting.pdf

Gentzkow, M., and Shapiro, J. M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 35–71. doi: 10.3982/ECTA7195

Goff, P. A., Eberhardt, J. L., Williams, M. J., and Jackson, M. C. (2008). Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. *J. Pers. Soc. Psychol*. 94:292. doi: 10.1037/0022-3514.94.2.292

Graf, S., Bilewicz, M., Finell, E., and Geschke, D. (2013). Nouns cut slices: effects of linguistic forms on intergroup bias. *J. Lang. Soc. Psychol*. 32, 62–83. doi: 10.1177/0261927X12463209

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol*. 96:1029. doi: 10.1037/a0015141

Greene, S., and Resnik, P. (2009). "More than words: syntactic packaging and implicit sentiment," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Boulder, CO: Association for Computational Linguistics), 503–511. doi: 10.3115/1620754.1620827

Haidt, J., and Graham, J. (2007). When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Soc. Justice Res*. 20, 98–116. doi: 10.1007/s11211-007-0034-z

Hall, D. G., and Moore, C. E. (1997). Red bluebirds and black greenflies: preschoolers' understanding of the semantics of adjectives and count nouns. *J. Exp. Child Psychol*. 67, 236–267. doi: 10.1006/jecp.1997.2404

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv[Preprint].arXiv:*1605.09096. doi: 10.18653/v1/P16-1141

Harris, L. T., and Fiske, S. T. (2015). Dehumanized perception. *Z. Psychol*. 219, 175–181. doi: 10.1027/2151-2604/a000065

Haslam, N. (2006). Dehumanization: an integrative review. *Pers. Soc. Psychol. Rev*. 10, 252–264. doi: 10.1207/s15327957pspr1003_4

Haslam, N., and Stratemeyer, M. (2016). Recent research on dehumanization. *Curr. Opin. Psychol*. 11, 25–29. doi: 10.1016/j.copsyc.2016.03.009

Hinds, J. (1977). Paragraph structure and pronominalization. *Paper Linguist*. 10, 77–99. doi: 10.1080/08351819709370440

Hodson, G., and Costello, K. (2007). Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychol. Sci*. 18, 691–698. doi: 10.1111/j.1467-9280.2007.01962.x

Katajamaki, H., and Koskela, M. (2006). "The rhetorical structure of editorials in english, swedish and finnish business newspapers," in *Teoksessa Proceedings of the 5th International Aelfe Conference* (Zaragoza), 215–19.

Kiritchenko, S., and Mohammad, S. (2018). "Examining gender and race bias in two hundred sentiment analysis systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (New Orleans, LA), 43–53. doi: 10.18653/v1/S18-2005

Kteily, N., Bruneau, E., Waytz, A., and Cotterill, S. (2015). The ascent of man: theoretical and empirical evidence for blatant dehumanization. *J. Pers. Soc. Psychol*. 109:901. doi: 10.1037/pspp0000048

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). "Statistically significant detection of linguistic change," in *Proceedings of the 24th International Conference on World Wide Web* (Florence: International World Wide Web Conferences Steering Committee), 625–635. doi: 10.1145/2736277.2741627

Levy, O., and Goldberg, Y. (2014). "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Baltimore, MD), 302–308. doi: 10.3115/v1/P14-2050

Manzini, T., Lim, Y. C., Tsvetkov, Y., and Black, A. W. (2019). Black is to criminal as caucasian is to police: detecting and removing multiclass bias in word embeddings. *arXiv[Preprint].arXiv:* 1904.04047. doi: 10.18653/v1/N19-1062

Marshall, S. R., and Shapiro, J. R. (2018). When "scurry" vs."hurry" makes the difference: vermin metaphors, disgust, and anti-immigrant attitudes. *J. Soc. Issues* 74, 774–789. doi: 10.1111/josi.12298

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," In *Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol.2* (Lake Tahoe, NV), 3111–3119.

Mohammad, S. M. (2018). "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)* (Melbourne, VIC). doi: 10.18653/v1/P18-1017

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin'words: lexical feature selection and evaluation for identifying the content of political conflict. *Polit. Anal*. 16, 372–403. doi: 10.1093/pan/mpn018

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). "Quotus: the structure of political media coverage as revealed by quoting patterns," in *Proceedings of the 24th International Conference on World Wide Web* (Florence: International World Wide Web Conferences Steering Committee), 798–808. doi: 10.1145/2736277.2741688

Nussbaum, M. C. (1999). *Sex and Social Justice*. Oxford: Oxford University Press.

Opotow, S. (1990). Moral exclusion and injustice: an introduction. *J. Soc. Issues* 46, 1–20. doi: 10.1111/j.1540-4560.1990.tb00268.x

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Number 47. Chicago, IL: University of Illinois Press.

Ott, B. L., and Aoki, E. (2002). The politics of negotiating public tragedy: media framing of the matthew shepard murder. *Rhetor. Public Affairs* 5, 483–505. doi: 10.1353/rap.2002.0060

Palmer, A., Robinson, M., and Phillips, K. K. (2017). "Illegal is not a noun: linguistic form for detection of pejorative nominalizations," in *Proceedings of the First Workshop on Abusive Language Online* (Vancouver, CA), 91–100. doi: 10.18653/v1/W17-3014

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count: Liwc 2001*. Mahway, NJ: Lawrence Erlbaum Associates.

Peters, J. W. (2014). *The Declineand Fall of the 'h' Word*, New York: New York Times.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). "Deep contextualized word representations," in *Proceedings of NAACL* (New Orleans, LA). doi: 10.18653/v1/N18-1202

Pew Research Center (2017). *Changing Attitudes on Gay Marriage*. Available online at: http://www.pewforum.org/fact-sheet/changing-attitudes-on-gay-marriage/

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., and Yang, D. (2019). Automatically neutralizing subjective bias in text. *arXiv[Preprint].arXiv:* 1911.09709. doi: 10.1609/aaai.v34i01.5385

Rashkin, H., Singh, S., and Choi, Y. (2016). "Connotation frames: a data-driven investigation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin), 311–321. doi: 10.18653/v1/P16-1030

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). "Linguistic models for analyzing and detecting biased language," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Sofia), 1650–1659.

Rothbart, M., and Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds. *Language, interaction and social cognition.* London: Sage.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). "Gender bias in coreference resolution," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (New Orleans, LA), 8–14. doi: 10.18653/v1/N18-2002

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol*. 39:1161. doi: 10.1037/h0077714

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2019). Social bias frames: reasoning about social and power implications of language. *arXiv[Preprint].arXiv:* 1911.03891.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., and Choi, Y. (2017). "Connotation frames of power and agency in modern films," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen), 2329–2334. doi: 10.18653/v1/D17-1247

Schmidt, A., and Wiegand, M. (2017). "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (Valencia), 1–10. doi: 10.18653/v1/W17-1101

Shah, N. (2019). "Asians are good at math" is not a compliment: stem success as a threat to personhood. *Harvard Educ. Rev*. 89, 661–686. doi: 10.17763/1943-5045-89.4.661

Sherman, G. D., and Haidt, J. (2011). Cuteness and disgust: the humanizing and dehumanizing effects of emotion. *Emot. Rev*. 3, 245–251. doi: 10.1177/1754073911402396

Shuman, E. L. (1894). *Steps Into Journalism: Helps and Hints for Young Writers*. Evanston, IL: Correspondence School of Journalism.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). "Analyzing the targets of hate in online social media," in *Tenth International AAAI Conference on Web and Social Media* (Cologne: AAAI), 687–690.

Smith, B. A., Murib, Z., Motta, M., Callaghan, T. H., and Theys, M. (2017). "Gay" or "homosexual"? The implications of social category labels for the structure of mass attitudes. *Am. Polit. Res*. 46:1532673X17706560. doi: 10.1177/1532673X17706560

Soller, K. (2018). *Six Times Journalists on the Paper's History of Covering Aids and Gay Issues*, New York: New York Times.

Steuter, E., and Wills, D. (2010). 'The vermin have struck again': dehumanizing the enemy in post 9/11 media representations. *Media War Conflict* 3, 152–167. doi: 10.1177/1750635210360082

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., et al. (2019). "Mitigating gender bias in natural language processing: literature review," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence), 1630–1640. doi: 10.18653/v1/P19-1159

Tipler, C., and Ruscher, J. B. (2014). Agency's role in dehumanization: non-human metaphors of out-groups. *Soc. Pers. Psychol. Compass* 8, 214–228. doi: 10.1111/spc3.12100

Tsur, O., Calacci, D., and Lazer, D. (2015). "A frame of mind: Using statistical models for detection of framing and agenda setting campaigns," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing), 1629–1638. doi: 10.3115/v1/P15-1157

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). "Metaphor detection with cross-lingual model transfer," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Baltimore, MD), 248–258. doi: 10.3115/v1/P14-1024

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., et al. (2020). Using social and behavioural science to support covid-19 pandemic response. *Nat. Hum. Behav*. 4, 460–471. doi: 10.1038/s41562-020-0884-z

Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., et al. (2020). Detecting east asian prejudice on social media. *arXiv[Preprint].arXiv:* 2005.03909.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., et al. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proc. Natl. Acad. Sci. U.S.A*. 114, 6521–6526. doi: 10.1073/pnas.1702413114

Wang, Z., and Potts, C. (2019). "Talkdown: a corpus for condescension detection in context," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong), 3702–3710. doi: 10.18653/v1/D19-1385

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Comput. Linguist*. 30, 277–308. doi: 10.1162/0891201041850885

Wierzbicka, A. (1986). What's in a noun? (or: how do nouns differ in meaning from adjectives?). *Stud. Lang.* 10, 353–389. doi: 10.1075/sl.10.2.05wie

Xu, C., and Liu, M. Y. (2020). *Social Cost With No Political Gain: The "Chinese Virus" Effect*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). "Gender bias in coreference resolution: evaluation and debiasing methods," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.2* (New Orleans, LA). doi: 10.18653/v1/N18-2003

Ziems, C., He, B., Soni, S., and Kumar, S. (2020). Racism is a virus: anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv[Preprint].arXiv:* 2005.12423.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 12 August 2020
doi: 10.3389/frai.2020.00058

Comparing Constraints on Contraction Using Bayesian Regression Modeling

*Laurel MacKenzie**

Department of Linguistics, New York University, New York, NY, United States

This paper has three goals: (1) to document the factors shaping *is*-contraction in Mainstream American English; (2) to assess the extent to which these factors also shape contraction of *has*; (3) to use shared patterns of contraction across the two verbs to draw conclusions about how the varying forms are represented grammatically. While *is* has two distinct phonological forms in variation, *has* has three. This necessitates regression modeling which can handle non-binary response variables; I use Bayesian Markov chain Monte Carlo modeling. Through this modeling, I (1) uncover a number of novel predictors shaping contraction of *is*, and (2) demonstrate that many of the patterns shown by *is* are also in evidence for *has*. I also (3) argue that modeling *has*-variation as the product of two stages of binary choices—a common treatment of three-way variation in variationist sociolinguistics—cannot adequately explain the quantitative patterns, which are only compatible with a grammatical model under which three distinct forms vary with each other. The findings have theoretical and methodological consequences for sociolinguistic work on ternary variables.

Keywords: contraction, English, copula, linguistic variable, Bayesian modeling, multinomial regression

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Joshua Waxman,
Yeshiva University, United States
Peter W. Culicover,
The Ohio State University,
United States

Correspondence:
Laurel MacKenzie
laurel.mackenzie@nyu.edu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *18 May 2020*
Accepted: *06 July 2020*
Published: *12 August 2020*

Citation:
MacKenzie L (2020) Comparing Constraints on Contraction Using Bayesian Regression Modeling. Front. Artif. Intell. 3:58. doi: 10.3389/frai.2020.00058

1. INTRODUCTION

Several English verbs can surface in at least two forms: one with all phonological material intact (e.g., [ɪz]), and one which is phonologically reduced and cliticized to its host (e.g., [z]). The variation between these two forms, known as contraction, has been investigated in a wide variety of corpora of both spoken and written language. This research has primarily focused on identifying the factors which condition one case of contraction in particular: the contraction of tensed forms of *be*, predominantly *is*.

Early work sought to identify the relative contributions of phonological, syntactic, and sociostylistic factors in the contraction of tensed *be*; later work has brought usage-based factors, such as predictability and persistence, into the picture. One particularly influential strand of work has compared the patterns of tensed *be* contraction in Mainstream American English to the patterns shown by tensed *be* absence in African American Vernacular English, and used the parallels between them to conclude that comparable processes drive the variation in both varieties. Studies of the constraints on contraction have shed light on broad theoretical questions about the nature of grammar; these include whether and how morphosyntax and phonology can interact (Anttila, 2016), and the extent to which grammar incorporates usage-based and processing constraints (Spencer, 2014).

Contraction of tensed verbs other than *be* has been less often examined. However, comparing patterns of contraction across different verbs can fill in our picture of how English verbs vary, and can answer questions about the generalizability of certain patterns which have thus far been attested only for *be*. This is turn speaks to questions of how variable phenomena are represented in the grammar, questions which have been addressed using contraction data since Labov (1969).

This paper contributes to developing a comprehensive understanding of tensed verb contraction in Mainstream American English (MAE) beyond the patterns evinced by tensed *be* alone. First, I pull together a variety of findings, not all of which have yet been considered together, on the factors that condition *is*-contraction. I examine their patterning on the largest data set of post-noun phrase *is*-contraction in spoken MAE to date.

Then, I explore the parallels between contraction of *is* and another verb which contracts in a similar way: *has*. Both verbs can surface in a non-syllabic form, represented in orthography as *'s*: an alveolar sibilant which agrees in voicing with the preceding segment. But *has* differs from *is* in a crucial way: when produced in spoken MAE, *has* has **two** possible syllabic realizations, one with an onset ([həz]) and one without ([əz]). Ternary variation like this complicates both variationist theory—where sociolinguistic variables are commonly represented as binary choices, even when this requires conflating two surface forms and opposing them to the third—and variationist methods—with traditional logistic regression analysis accommodating no more than two possible outcomes. For this reason, no study to date has yet adequately analyzed the quantitative patterning of *has* in spoken language in a way that recognizes its three unique surface realizations **and** allows all three possible forms to vary independently. The present study does this, capitalizing on a recent movement employing Bayesian multinomial modeling for the analysis of ternary linguistic variation (e.g., Levshina, 2016; Grafmiller et al., 2018; Dilley et al., 2019).

Doing this allows me to address an additional issue: the underlying representation of the three different forms of *has*. In previous work (MacKenzie, 2013), I argued in favor of a treatment under which the tripartite variation in *has* was best understood as deriving from a cascade of two binary choices. The three forms on the surface mapped onto two forms underlyingly; the third form was derivative from one of the two selected in the first-stage choice. This, I argued, explained certain quantitative patterns in the data. When this analysis was assumed, the rate at which different forms of *has* were used in different contexts paralleled the rates for the analogous forms of *is*, suggesting unity of the two contraction processes at an abstract level. That analysis, though, considered only two contextual effects on the contraction of *is* and *has*. The present study considers thirteen. This not only expands the testing ground on which to look for analogous behavior in the patterning of the two verbs; it deepens our understanding of the contextual factors that affect variation in these two verbs in the first place.

In this paper, I answer the following questions:

1. Which factors condition the alternation between contracted and full forms of *is*?
2. Do those same factors condition the alternation between the contracted form of *has* and its two other possible forms?
3. Does the patterning of *has* lend support to the analysis under which the three surface forms are derived from two stages of binary choice?

These questions echo Labov's (1969, p. 760) research program of identifying "the most general form of linguistic rule" when similar-seeming patterns recur across different variable phenomena. Shared patterns of variation can be taken as evidence for structural unity of varying items in speakers' mental grammars. In fact, I do find that there are a number of shared patterns between contraction of *is* and contraction of *has*. But I also find that the *has* patterns cannot be fully accommodated under the earlier two-stage analysis. The forms of *has* appear to vary in a ternary way, complicating our understanding of what a linguistic variable can look like.

The findings have relevance not only for studies of tensed verb contraction in English, but for longstanding questions of the nature and representation of variable phenomena. They additionally carry methodological importance for variationist sociolinguistics. It is not uncommon for researchers, when faced with an alternation that is more than binary, to group variants together for regression analysis. However, a longstanding theoretical tenet of variationist research is that regression models are meant to represent grammar (Cedergren and Sankoff, 1974). Grouping two variants together and opposing them to a third implies that, at some level, the speaker makes such a choice. Grouping as a methodological exigency thus has theoretical consequences that may be unwarranted. In the present paper, I demonstrate that allowing three variants to vary independently in a statistical model can shed light on the relationship between those variants without the researcher having to impose any such relationship on the analysis. Accordingly, the paper includes a call to action for variationist sociolinguists who work with non-binary variables to branch out into different modeling techniques.

2. BACKGROUND

2.1. Analyzing Contraction

The variation under study in this paper is the phonological realization of two tensed English verbs: *is* and *has*. These verbs can variably surface in a number of phonologically distinct forms. The verb *is* has two distinct forms in which it can surface: a single-consonant form (which agrees in voicing with the preceding segment), and a syllabic form, as shown in 1[1]. The following examples are taken from the Switchboard Corpus (Godfrey and Holliman, 1997; see section 3.1 for more on the corpus).

(1) Forms of *is*.
Yeah, Salzburg'[z] nice. Austria'[z] nice. Europe [əz] nice!

[1] The level of stress and reduction of the vowel in the syllabic form also varies, but it is standard practice in the variationist literature on contraction to abstract away from this (Labov, 1969) or to study it as a separate phenomenon, independent of contraction (Spencer, 2014).

TABLE 1 | Forms of *is* and *has*.

Name	Description	Example
Contracted form	Single-consonant form	[z] "is," "has"
Full form	Phonologically intact form	[əz] "is," [həz] "has"
Intermediate form	Third form (*has* only)	[əz] "has"

(sw1151)[2]
[z]: single-consonant form
[əz]: syllabic form

The verb *has* has three distinct forms in which it can surface: a single-consonant form, a syllabic form with no onset, and a syllabic form with an onset, as shown in 2.

(2) Forms of *has*.
This spring [həz] been a little hard to keep up the pace because we've had a lot of family activities: my wife [əz] taken up skiing [...] she[z] taken up snow skiing. (sw1402)
[z]: single-consonant form
[əz]: syllabic form, no onset
[həz]: syllabic form with onset

In other words, *is* and *has* both show variation between a single-consonant form ([z]) and a form with all phonological material intact (for *is*, [əz]; for *has*, [həz]). But *has* differs from *is* in additionally allowing a third variant, the syllabic, onsetless [əz].

For ease of exposition, I give the three different forms unique names (**Table 1**). I follow the literature in using the term "contraction" to refer to the alternation between the contracted form and any other form(s) of a given verb.

In previous work (MacKenzie, 2013), I followed a sizable body of work in phonology and morphology and analyzed *is*-contraction as variable allomorphy. In other words, when producing the third singular present form of *be*, a speaker has a choice between two options: one that consists of a single consonant (the contracted form), and one with all phonological material intact (the full form).

This then raises the question of what kind of choice a speaker has when going to produce the third singular present form of auxiliary *have*. MacKenzie (2013) considers two possibilities. The first, a ternary analysis, treats all three surface forms of *has* as represented underlyingly. Variation in *has* realization under this approach would consist of variable three-way allomorphy: a choice between full, contracted, and intermediate forms. The second possibility consists of two binary choices: first, a choice between the contracted form and the full form, as for *is*; then, in cases where the full form has been selected, a second choice between producing the full form as-is or reducing it to the intermediate form via /h/-deletion, an independently attested fast-speech reduction phenomenon in English. This second possibility brings contraction of *has* more in line with contraction of *is*. It also suggests that the choice between full and contracted forms, which takes place in a similar way underlyingly for the two verbs, may pattern in the same way on the surface. Indeed, this is what I found in MacKenzie (2013). Contracted forms of *is* and *has* are selected at very similar rates for the two verbs. Additionally, contracted forms of *is* and *has* both show identical effects of being dispreferred after longer noun phrase hosts. By contrast, the choice between intermediate and full forms of *has* shows no sensitivity to host phrase length. This suggests that the factor of host phrase length operates on a choice between contracted and full, and not on a later choice point that may occur between full and intermediate.

Still unresolved is whether the other factors that condition contracted forms of *is*—which will be detailed in the following section and confirmed in the first set of results presented in this paper—operate in the same way on contracted, but not intermediate or full, forms of *has*. If they do, this will constitute even more evidence in favor of the two-stage analysis of *has* presented in MacKenzie (2013).

To set the stage for this analysis, I survey the existing work on contraction of *is* in the next subsection.

2.2. Variation in Tensed *Be*

The bulk of quantitative corpus research on contraction addresses variation in tensed *be*, and within that, there is considerable research on contraction of the third singular form *is*. Despite some differences in the data used across different studies—spoken vs. written language, sociolinguistic interviews vs. telephone conversations—several key factors are consistently found to shape the alternation between *is* and *'s*. Many of these have to do with the nature of what I will call the "host phrase" of the contractable verb: the phrase onto which the verb cliticizes when it contracts[3]. Also relevant are properties of what I call the "host word"—the word immediately preceding the contractable verb—and the verb's complement—the constituent following the contractable verb.

One of the strongest effects on contraction of *is* is whether the verb's host phrase is a pronoun or a non-pronominal noun phrase (henceforth "NP"). Speakers use the contracted form of *is* at near-categorical rates after pronominal host phrases, and much lower rates after NP host phrases (Labov, 1969; Rickford et al., 1991; McElhinny, 1993; MacKenzie, 2013; Barth and Kapatsinski, 2014; Spencer, 2014; Bresnan, 2018). Due to this near-invariance, some researchers have opted to analyze *is*-contraction after NP host phrases separately from pronominal host phrases, or even dispense with post-pronominal data altogether, because the contraction rate is so high. I adopt the latter approach in the present study, examining contraction only after NP host phrases. In section 3.2, I describe the exclusion criteria I used to achieve this.

Another strong effect on *is*-contraction is the length of an NP host phrase. Even with pronominal host phrases excluded, longer

[2] Numbers in parentheses are speaker identification numbers from the Switchboard corpus. To facilitate readability, disfluencies and repetitions have been removed from example transcripts, and capitalization and punctuation have been added.

[3] This is typically the verb's subject, but in a *wh*-question (e.g., *How old's your son?*) it may be a different constituent.

host phrases disfavor contraction (MacKenzie, 2013; Spencer, 2014; Bresnan, 2018). Host phrase length can be operationalized in a number of different ways; there is some evidence that orthographic word count predicts the variation somewhat better than other measures (MacKenzie, 2012, chapter 5).

Semantic and phonological characteristics of the host phrase also play a role in conditioning the variation. *Is*-contraction has been found to be sensitive to host phrase animacy, with more contraction after human-referent than non-human referent host phrases, an effect that is not reducible to a confound with proper noun status (McLaughlin, 2014). And studies generally find *is* to contract more after an immediately-preceding vowel than a consonant, and more after a voiced consonant than a voiceless one (Labov, 1969; MacKenzie, 2012; Spencer, 2014). Though surrounding syllable stress has been hypothesized to play a role in contraction (Anttila, 2017), it has not been found to affect *is*-contraction in the two studies that have looked (MacKenzie, 2012; Bresnan, 2018).

Additionally, a widely-discussed effect on *is*-contraction is the syntax of the verb's complement. The sociolinguistic literature on *is*-variation in African American Vernacular English (AAVE)—which allows both *is*-contraction and *is*-deletion—has tended to differentiate five complement types: nominal, locative, adjectival, progressive verb, and future *gonna/going to*. Deletion of *is* in AAVE shows clear sensitivity to this factor, with the ordering of environments given in the previous sentence reflecting a commonly replicated hierarchy from least deletion-favoring to most deletion-favoring (Sharma and Rickford, 2009). However, contraction of *is*, in both AAVE and MAE, has shown less clear patterning, and it is difficult to compare across studies which have operationalized this factor in different ways. That said, there is a general trend by which verbal complements—progressive verbs and/or futures—favor contraction more than other complements (Labov, 1969; McElhinny, 1993; Barth and Kapatsinski, 2014; Spencer, 2014; MacKenzie, 2016).

Well-studied in recent literature are measures of the predictability, or conditional probability, of the contractable verb given surrounding words. Some researchers have found that *is*-contraction is more likely when the verb is highly probable given one or two surrounding words (Frank and Jaeger, 2008; Barth and Kapatsinski, 2014; Spencer, 2014), though the results depend on whether pronominal subjects are included in the analysis or not. There is also evidence that *is*-contraction displays persistence, that is, that speakers show a tendency to reuse whichever variant of *is* was previously used (Barth and Kapatsinski, 2014; Spencer, 2014; Bresnan, 2018). Though a few studies have considered the effect of speaking rate on *is*-contraction (Frank and Jaeger, 2008; MacKenzie, 2012; Spencer, 2014), it does not show a convincing, statistically significant effect in any of them.

Finally, where sociodemographic factors are concerned, there is some evidence that *is*-contraction shows effects of speaker sex/gender—with speakers identified by the corpus as male using contraction more than those identified as female (MacKenzie, 2012)—and speaker age, with younger speakers contracting more than older ones (Rickford et al., 1991; MacKenzie, 2012; Bresnan, 2018). At the same time, there is no evidence that speakers style-shift *is*-contraction in speech (Finegan and Biber, 1986; McElhinny, 1993; MacKenzie, 2012).

2.3. Variation in Other Tensed Verbs

There is much less research on the variable phonological realization of other tensed verbs, including auxiliary *has*, the other verb analyzed in this paper. Where *has* has been examined, researchers have generally opposed the single-consonant "contracted" form ([z] or [s] depending on voicing of the preceding segment) to the other forms ("intermediate" [əz], "full" [həz]) (McElhinny, 1993; Frank and Jaeger, 2008). This seems to presume a particular analysis of variant choice—that speakers make a binary choice between the contracted form and the other two combined—though this is not made explicit. The results are also difficult to generalize over, due to small token counts (McElhinny, 1993 examines only 76 tokens of *has*) and to researchers collapsing across forms (Frank and Jaeger, 2008 analyze *has, have,* and *had* together). Nevertheless, we can glean some patterns. *Has*-contraction shows the same favoring effect of a pronoun (as opposed to an NP) host phrase as *is*-contraction, and, among NP host phrases, the same effect of host phrase length in words (McElhinny, 1993; Frank and Jaeger, 2008; MacKenzie, 2012). Frank and Jaeger additionally find an effect of verb predictability, in keeping with that found for *is*-contraction. Analyses of preceding segment, speaking rate, and speaker sociodemographic factors are inconclusive, with some of the aforementioned three studies finding them, and others not.

2.4. Current Contribution

As the previous subsection emphasized, the present paper is virtually unique in analyzing contraction of *has* alongside the very similar contraction of *is*. Research that has compared these two verbs has not operationalized the forms of *has* as I do here, i.e., as three distinct forms that may vary independently.

In addition, the present paper expands our body of knowledge on the contraction of *is*. Though much research on *is*-contraction exists, the present study improves upon previous studies in two key ways. First, the present paper uses auditory coding of the variation, rather than relying on transcripts, which may not accurately reflect spoken language. Second, compared to other studies of *is*-contraction that do make use of auditorily-coded data, the present paper employs a much larger data base. Even though the data has been restricted to only those tokens of *is* with NP host phrases, my data base of 5,642 tokens is four times as large as that of Bresnan (2018), and nearly 10 times as large as that of Spencer (2014). This allows for increased statistical power, and uncovers novel results.

Finally, I see the present paper as making important methodological and theoretical contributions where the study of non-binary variation is concerned. Variables with more than two variants have long posed a problem for sociolinguistic research, for reasons of method—logistic regression models require outcome variables to be binary—and for reasons of theory—the original conception of the variable rule was that a single input variably yielded a single output (Cedergren and Sankoff, 1974; Wolfram, 1991). To get around these problems, researchers have resorted to strategically grouping variants together. So, in cases of ternary variation, researchers will combine data from two variants and oppose them to the third: see, for instance, a large literature on /t/ variation in regional British Englishes, where attested forms of /t/ include [t], [tʔ], and [ʔ], and various grouping strategies are employed (e.g., Foulkes et al., 2005; Straw and Patrick, 2007; Drummond, 2011). But it is not often acknowledged that grouping variants carries implicit theoretical assumptions about the structure of variation. A longstanding theoretical tenet of variationist research is that regression models are meant to represent grammar (Cedergren and Sankoff, 1974). While the earliest work to group non-binary variables explicitly linked the grouping procedure to a particular theoretical treatment of the variation (Labov, 1969), many more recent studies that group don't recognize the tacit grammatical claims that their grouping implies.

Another problem with grouping is that it can present a misleading picture of the constraints on variation. This was notably pointed out by Rickford et al. (1991) in their critique of Labov's (1969) study of copula contraction and deletion in AAVE. Labov defined copula "contraction" by opposing contracted and deleted forms of the copula to full forms, because, the theory went, all deletions had contracted forms at some point in their derivational history. He defined copula "deletion" by opposing deleted to contracted forms, omitting full. He then demonstrated that contraction and deletion were conditioned in the same way, which he argued supported an analysis under which contracted and deleted forms shared an underlying representation, and hence justified his grouping. But, as Rickford et al. pointed out, when contraction is calculated by grouping together contracted and deleted forms, it will inevitably be influenced by the patterns of deletion, particularly in cases where deleted forms greatly outnumber contracted ones.

I suggest that, when faced with non-binary variation like this, multinomial regression modeling is an important alternative to grouping, both in cases where a researcher does not want to impose a particular theoretical analysis on the data (such as AAVE copula variation, see McLaughlin, 2014), and in cases where there is no immediately obvious two-stage analysis to be imposed (such as English ternary genitive variation, see Szmrecsanyi et al., 2016). As an additional point in its favor, multinomial regression modeling has been found to explain variation as well as models that assume two stages of binary choice, at least for some variables (Sankoff and Rousseau, 1989).

In this paper, I take the multinomial model of *has*, under which all three forms are allowed to vary independently, as a null hypothesis. Then, I compare the factors that condition speakers' choices between contracted and the other two forms of *has* to the factors that condition speakers' choice between contracted and full forms of *is*. If contracted forms of *has* behave in opposition to the other two forms, and they are conditioned in similar ways to contracted forms of *is*, we have evidence to support the analysis of *has* variation under which speakers make a first-stage choice between contracted and full—just as they do for *is*—and then, where applicable, a second-stage choice between full and intermediate.

3. METHODS

3.1. The Data

The data for the present study come from the Switchboard Corpus (Godfrey and Holliman, 1997). Switchboard is a transcribed corpus of telephone conversations between 542 native speakers of American English, paired at random by a robotic operator and assigned a topic to elicit a 5- to 10-min conversation. The corpus was collected between 1991 and 1992, and consists of about 240 h of speech (roughly 3 million transcribed words) across approximately 2,400 conversations.

Data were collected as described in section 3.2, and coded for the predictors enumerated in section 3.3. Data points with an NA value for any of the predictors of interest were omitted from analysis. This resulted in 5,642 tokens of *is* and 699 tokens of *has*.

3.2. Defining the Variable Context and Data Extraction

As mentioned in section 2.2, data for the present study were restricted to only those tokens of *is* and *has* following non-pronoun subjects. To this end, it was important to identify what counted as a pronoun. Data was excluded from the present study if the host phrase was any of the following: a personal pronoun (e.g., *she, he*), an expletive pronoun (*there, it*), a *wh*-pronoun (e.g., *what, who, where, whatever*), a demonstrative pronoun (e.g., *that*), an indefinite pronoun (e.g., *everybody, someone, anything, one*), a possessive pronoun (e.g., *mine*), or the locative *here*. This is a more conservative definition of what counts as a "pronoun host phrase" than others have used, but it is justified by the finding that pronoun-like host phrases, such as indefinite pronouns, have significantly higher rates of contraction than single-word NP hosts (MacKenzie, 2012). This suggests that contraction shows special behavior when the host phrase is a closed-class lexical item; for that reason, I omit any data points where the verb's host phrase is anything pronoun-like. Though this removes a relevant factor in the choice of contracted form, we are still left with a number of other factors to examine.

The first step in obtaining data was to search the corpus for the variants of each verb when occurring in non-post-pronominal contexts. This was done using a Python script. The script searched for tokens of the targeted verbs whose immediately preceding word did not fall into the category of pronouns listed above. To filter out tokens of main verb *has*, which does not contract in American English (Hughes et al., 2012, p. 23), the script returned hits for *has* and *'s* only when followed by a past participle with no more than three words optionally intervening. Past participles were defined as any word ending in *-en* or *-ed*, or on a list of 129 irregular past participles (e.g., *begun, gone*). All instances of *has* were scrutinized, and tokens of main verb *has* that slipped through were removed from the data.

To target forms of *is*, the script searched for *is* and *'s*. All instances of *'s*, which can reflect several different morphemes in English, were scrutinized. Instances of *'s* that were actually contracted forms of *has* were retained in the data only when they had not been picked up by the previous search. Instances of *'s* that were actually the possessive morpheme were removed from the data.

After this initial stage of data collection, the second step was to eliminate data points where *is* or *has* occurred in an environment where contraction is blocked. This follows traditional variationist methodology, and ensures that the analyst only studies those environments in which each form of the dependent variable is grammatical, preventing results from being skewed (Labov, 1972; Tagliamonte, 2006). The numerous environments where the full range of variants of *is* and *has* is prevented from surfacing can be found in MacKenzie (2012), chapter 3, and references cited therein.

Tokens were also omitted from study when they contained a negated verb, since three variants are possible there (e.g., *is not*, *'s not*, or *isn't*). Finally, a single Switchboard speaker was observed to use copula deletion; tokens from this speaker were omitted, since the availability of this third variant of *is* skews the distribution of forms relative to other speakers.

Once the data had been obtained, the author listened to each instance of *is/has* in the data and coded each occurrence of the dependent variable as contracted, full, or (*has* only) intermediate. As part of this stage of data extraction, tokens were excluded if the verb was contrastively stressed or if the speaker paused between the host word and the verb (MacKenzie, 2012).

Even though contraction of *is* and *has* is represented orthographically in English, it was important to listen to each instance of the dependent variable and code it auditorily, rather than relying on Switchboard transcriptions. There were two reasons for this. First, there is no standardized way of representing the intermediate form of *has* ([əz]) orthographically, but it is still a phonologically distinct variant and should be coded as such. In fact, the vast majority of tokens identified by the author as phonologically intermediate were orthographically represented in the transcriptions as full forms (179 out of 182). Second, there is reason to believe that Switchboard's transcriptions of contracted and full forms are not reliable. According to the Switchboard manual, transcribers were "always permitted to spell out forms in full, even if the pronunciation suggests the contracted form." Indeed, of the forms of *has* and *is* that were identified by the author as contracted, a sizable proportion of them were found to have been transcribed as full by Switchboard's transcribers (*has*: 21%; *is*: 35%). (Forms identified by the author as full were indeed transcribed as such, at a rate of 100% for *has* and 99% for *is*). For this reason, auditory coding of the dependent variable was essential. The author carried out all such coding.

Finally, each data point was coded by the author or a trained research assistant for a number of predictors, described in the following subsection.

3.3. Predictors

Modeling included three random intercepts: speaker, word preceding the target, word following the target. (Here and henceforth, "target" refers to the contractable verb.) Speakers and words with five or fewer observations in the data were recoded as "other" following Levshina (2016, p. 253). The fixed-effect predictors coded for were:

- **Host phrase length in orthographic words**: a continuous measure.
- **Host phrase humanness**: categorical, treatment coded, with three levels, following Rosenbach (2005) and Wolk et al. (2013): **human** (default); **collective**, comprising organizations and "temporally stable groups of humans with potentially variable concord" (Wolk et al., 2013, 396); **non-human**.
- **Host phrase proper nounhood**: categorical, treatment coded, with two levels, **no** (default) and **yes**.
- **Preceding segment**: categorical, sum coded, with levels **voiced consonant, voiceless consonant, high vowel, other vowel, R**. Segments were identified based on a transcription of the preceding word taken from the CMU Dictionary v.0.7 (Weide, 2008). Words not in the dictionary were transcribed by hand. The subdivision of consonants by voicing is informed by previous findings (e.g., Spencer, 2014). The subdivision of vowels into high and other was based on my experience coding the dependent variable: I often had difficulty determining whether an instance of *is* following a high vowel was contracted or not[4]. /ɹ/-colored final vowels were given their own category due to uncertainty concerning whether they should be considered vowels or consonants.
- **Preceding and following syllable stress**: categorical, sum coded, with levels **monosyllabic, primary, secondary, unstressed**. Syllable stress was obtained based on the transcriptions provided in the CMU Dictionary v.0.7. Words not in the dictionary were transcribed by hand. Due to small Ns, the *primary* and *secondary* categories of following syllable stress were combined as *stressed* for *has*.
- **Complement syntax (*is* only)**: categorical, sum coded, with levels **unknown** (speaker changes direction or restarts), **noun phrase** (including gerunds), **determiner phrase, quantifier phrase**[5], ***wh*-phrase, past participle, adjective phrase, number phrase, prepositional phrase, locative prepositional phrase, progressive, future**. Cases where a disfluency and/or an adverb immediately followed the target were coded for the syntax of the constituent following the disfluency/adverb. This is a larger number of categories than has been identified in previous studies, but ambiguity in previous researchers' methods made it difficult to apply any previous coding scheme to the present data, so the decision was made to err on the side of caution and make more distinctions than were potentially necessary. Still, some issues remain with the coding: for instance, *about to* (as in *Summer's about to be here*) was coded as a prepositional phrase, but semantically, it has a future meaning. Ascertaining the behavior of such syntactically–semantically mismatching following constituents is an interesting direction for future work.

[4] Such "neutralization environments" are often omitted in variationist work, but some researchers advocate for their inclusion as long as the modeling can account for their potentially exceptional behavior (e.g., Tanner et al., 2017).

[5] This category included four tokens in which the complement of *is* is quotative *be like*, following Haddican and Zweig's (2012) analysis of quotative *be like* as taking a silent *something* quantifier phrase complement.

- **Speaker sex/gender**: categorical, sum coded, with levels **male, female** based on the information provided in the corpus.
- **Speaker year of birth**: continuous, centered around the median, rescaled to decades.
- **Previous form of verb**: categorical, treatment coded, with levels **none** (default), **full, contracted, intermediate** (for *has* only). This predictor checks for persistence, and compares the likelihood of contraction of tokens that follow a previous instance of the verb to the likelihood of contraction of the first token of a conversation. Coding for this predictor was done on a speaker-by-speaker basis, so cross-speaker persistence or accommodation effects were not allowed for. Instances of *has* where it was functioning as a main verb were not counted as previous forms of *has*. Also uncounted were instances of *'s* where it was functioning as a possessive marker or as a contracted form of the other verb (e.g., contracted *is* when the target was *has*, etc.). Future work can explore the possibility of persistence effects between phonologically identical but morphologically distinct forms like these (and see also Tamminga, 2016).
- **Relative speaking rate**: a continuous measure reflecting the ratio between the speaking rate of the annotation unit containing the target and the speaker's average speaking rate across the entire corpus. The higher the value, the faster the speech is, relative to the speaker's average.
- **Following disfluency**: categorical, treatment coded, with levels **no** (default) and **yes**, reflecting whether the word immediately following the target was *uh* or *um*.
- **Forward transitional probability**: a continuous measure reflecting the conditional probability of the target given the preceding word. Calculated as the corpus-internal frequency of the preceding word + target bigram divided by the frequency of the preceding word. Log-transformed.
- **Backward transitional probability:** a continuous measure reflecting the conditional probability of the target given the following word. Calculated as the corpus-internal frequency of the target + following word bigram divided by the frequency of the following word. Log-transformed.

Spearman correlations were used to check continuous predictors for collinearity. For both verbs, host phrase length and forward transitional probability were found to be weakly negatively correlated (rho $= -0.328$ for *is*, rho $= -0.318$ for *has*). This is unsurprising: longer subjects are more likely to be structurally complex, with embedded phrases causing them to end in items like verbs and prepositions, which are unlikely to be themselves followed by a(nother) verb. Accordingly, the log-transformed measure of forward transitional probability was residualized by host phrase length for each verb, and this residualized predictor was used in modeling.

3.4. Modeling

While *is*-contraction is easily modeled using the binomial (two-outcome) mixed-effects logistic regression models that have become commonplace in variationist research, the three-way variation shown by *has* is not. As Sankoff and Rousseau (1989, 6) observe, there are two fundamental approaches to modeling a three-variant variable like this: as a single choice between three options each time a speaker goes to produce a form, or as two binomial choices: first between one form and the other two combined, and then between those latter two forms. Like *is*-contraction, the second of these two options can be easily modeled with (two rounds of) binomial logistic regression, but at the downside of imposing a particular analysis on the data (see section 2.4).

For this reason, I analyze *has*-contraction with multinomial logistic regression, a simple extension of binomial logistic regression which allows the user to compare a reference or default category to each of the other possible outcomes (Levshina, 2015b). And, in order to accommodate the inclusion of random effects, which have been argued to be essential in sociolinguistic research (Johnson, 2009), I implement Bayesian modeling using R's `MCMCglmm` package (Hadfield, 2010a). For consistency, I use Bayesian modeling for both verbs: a binomial model for *is*, and a multinomial one for *has*. Recent linguistic papers that make use of multinomial MCMCglmms include Levshina (2016), Grafmiller et al. (2018), and Dilley et al. (2019). For a detailed description of the philosophy and methodology behind these models that is geared to a linguistic audience, the reader is pointed to the first two of these articles. Levshina (2015a) provides a brief tutorial, again for a linguistic audience, on getting started with MCMCglmm modeling; for a more detailed tutorial and primer on these models, consult Hadfield (2010a,b, 2019). Below, I briefly describe these models and summarize what distinguishes them from the logistic regression models that sociolinguists are used to.

MCMCglmm implements Bayesian Markov chain Monte Carlo modeling. The models used here are Bayesian in that they require the user to specify prior beliefs about the probability distributions of the model parameters; after considering the data, they output posterior probability distributions for each parameter of interest. The models also make use of Markov chain Monte Carlo methods to estimate the posterior probabilities, generating representative random values from these distributions and then approximating the posterior probability distributions from these values. The output of an MCMCglmm model can be interpreted like the output of a logistic regression model fit with `lme4` in R (R Core Team, 2017): coefficients in log odds are provided for the different levels of each categorical independent variable; these indicate the change in log odds associated with that level of using the non-default variant of the dependent variable. For continuous predictors, coefficients reflect the change in log odds of using the non-default variant of the dependent variable with each one-unit increase of the predictor. Positive coefficients reflect a change in log odds in favor of the non-default variant; negative coefficients reflect a change in log odds in favor of the default variant, or reference level. However, unlike in traditional logistic regression modeling, where a single value is estimated for each coefficient, in Bayesian modeling, coefficients reflect averages calculated over the probability distributions output by each of the many iterations of the model.

The models presented here contain two major departures from the frequentist binomial logistic regression models that sociolinguists are accustomed to. The first stems from their

Bayesian nature. Model coefficients are not reported with *p*-values to reflect the probability that the result evident in the data would hold were the null hypothesis true of the wider population. Instead, researchers report 95% Highest Posterior Density (HPD) intervals, or "credible" intervals: intervals in which 95% of the posterior probability density lies. If the 95% credible interval for a predictor does not include 0, we can be reasonably confident that the predictor of interest has a non-zero effect on the data, i.e., the probability of the coefficient for the predictor of interest being non-zero is 0.95. In this way, Bayesian models can be used to estimate the probability of a given parameter taking on a specific value. As Grafmiller et al. (2018) argue, the philosophy behind the Bayesian approach to statistical analysis—estimating the probability of a hypothesis given the data, rather than the probability of one's data given a (null) hypothesis—is intuitively easier to grasp than the traditional frequentist method. The second departure from traditional sociolinguistic modeling is seen in the output of the multinomial model for *has*. Because multinomial logistic regression compares the output of each non-default variant to the default variant, each predictor in a three-way multinomial model has a set of two coefficients, one for each of the non-default variants as compared to the default. Both are interpreted as in binomial logistic regression: again, a positive coefficient favors use of the variant in question over the default; a negative coefficient favors the default over the variant in question. It is possible, for instance, for both non-default predictors to have positive coefficients, indicating that both are favored over the default for a given level of an independent variable.

Running MCMCglmm models requires setting prior probabilities. Following the researchers cited at the beginning of this subsection, I used weakly informative priors defined following the specification for categorical distributions given in Hadfield (2010b, p. 21–24). Another aspect of MCMCglmm that must be set by the user is the number of iterations the model runs for. For the *is* data, I ran 60,000 iterations, sampling every 50th iteration, and discarding the first 10,000 to correct for initial sampling bias (the "burn-in" period). This left 1,000 posterior estimates of each parameter, from which averages and 95% credible intervals are calculated and presented in the following section. For *has*, where there is much less data, I ran 600,000 iterations, sampling every 250th, and discarding the first 50,000. This left 2,200 estimates. Both models were checked for convergence by using strategies to assess autocorrelation suggested in Levshina (2015a, 2016). Checking for autocorrelation (i.e., non-convergence) in each model using the `autocorr()` function in R as well as by visually examining trace plots of the model's parameters revealed that the model chains had mixed well. Model specifications are provided in the **Supplementary Material**.

Following Grafmiller et al. (2018), model accuracy was assessed by comparing predicted values generated by the model to observed values for each data point. This allows the construction of a confusion matrix and the calculation of accuracy rates (percent of predicted forms which were correct) and recall rates (percent of observed forms which were correctly predicted). For the binomial model, I also use these predicted values to calculate two measures of model predictive accuracy: Somers' D, which calculates the correlation between the observed values and the log odds of using the default variant for each data point, and the corresponding receiver operating characteristic curve area C.

For both verbs, *contracted* was taken as the default level of the response. If the two-stage analysis of *has*-contraction proposed by MacKenzie (2013) is correct, then we expect to see two choices patterning the same way: the choice between contracted and full forms of *has*, and the choice between contracted and intermediate forms of *has*. This is because the analysis posits a single stage of choice at which speakers decide between using a contracted form and using a full form, which itself may or may not eventually become an intermediate form. Accordingly, the environments in which speakers prefer a contracted form of *has* should equally be the environments in which speakers disprefer the other two forms of *has* (see McLaughlin, 2014 for a very similar approach to the contraction and deletion of *is* in AAVE). Additionally, the models can also answer the question of whether the factors that lead speakers to choose contracted forms of *has* are the same as those that lead speakers to choose contracted forms of *is*, again as suggested by MacKenzie (2013). This would lend further support to the two-stage analysis, opening up the possibility that contraction of *is* and *has* can be unified as a single abstract alternation between contracted and full, with intermediate forms being derivative, stemming from a later process.

To this end, in the next section, I first present the results from the *is* model, and then present the results of the *has* model, to answer the two questions of whether the same environments prefer contracted forms of both verbs, and whether those same environments equally disprefer the two non-contracted forms of *has*.

4. RESULTS

Before turning to the MCMCglmm outputs for each individual verb, it is instructive to consider the overall rates of variant use. **Figure 1** shows this. It is immediately apparent that contracted forms of each verb (represented by the orange [uppermost] sections of each bar) are used at an almost identical rate (*has*: 36.6%, *is*: 35.5%). When the two non-contracted forms of *has* are grouped together and opposed to the contracted form, a chi-square test finds no significant difference in distribution of forms between the two verbs ($\chi^2 = 0.241$, df = 1, $p = 0.623$). This replicates the finding from MacKenzie (2013), but with a considerably larger data set. It also constitutes a first piece of evidence in support of that earlier analysis, under which a first step of choice between contracted and other form(s) applies in a similar way across verbs. Subsequent evidence in favor of—and against—this analysis will be taken from the factors that condition speakers' choice of contracted vs. other forms, to be discussed on a verb-by-verb basis in the two subsections that follow.

4.1. *Is*

4.1.1. Model Predictions and Accuracy

The binomial model of *is*-contraction predicts verb form with a high degree of accuracy (C = 0.875, D = 0.749). **Table 2** shows the confusion matrix of predicted and observed forms. The model predicts the correct form 80% of the time; this is a

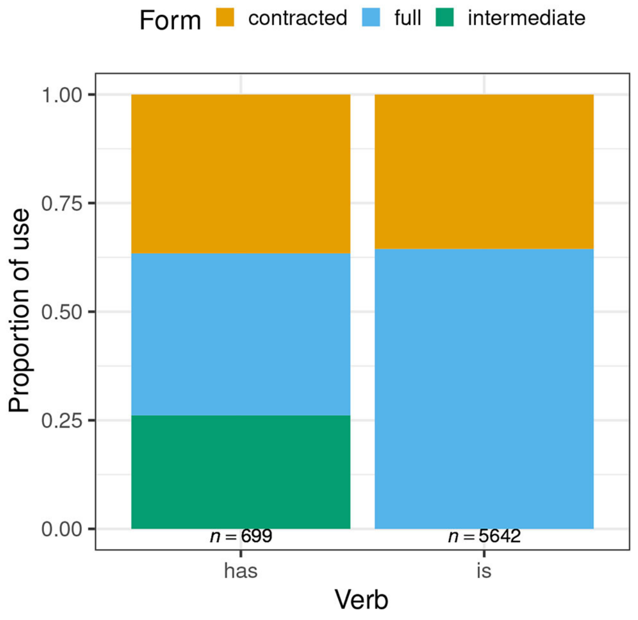


FIGURE 1 | Distribution of variants of post-NP *has* and *is*.

TABLE 2 | Confusion matrix for *is*-contraction.

	Predicted		
	Contracted	**Full**	**Total**
Observed			
Contracted	1,288	718	2,006
Full	423	3,213	3,636
Total	1,711	3,931	5,642

significant increase over simply guessing the most frequent form every time, which would give an accuracy rate of 64% ($p_{binom} < 0.001$). We can thus be confident that the model is a good fit for the data.

4.1.2. Results

Figure 2 provides a graphical representation of the estimated log odds of full form usage, along with 95% credible intervals, for each predictor in the *is* model. **Table 3** provides the same information—the posterior means and 95% Highest Posterior Density (credible) interval boundaries for each predictor—along with two probabilities, in the last two columns: the probability that the true coefficient for the predictor is less than 0 (i.e., that the predictor favors the use of the default—contracted—form), and the probability that the true coefficient for the predictor is greater than 0 (i.e., that the predictor disfavors use of the contracted form and favors use of the full form). This can help contextualize the results presented visually: even a predictor whose 95% credible intervals cross 0 may nonetheless be predicted with fairly high probability to influence the variation.

Four predictors pertaining to the syntax, semantics, and phonology of the host phrase have clear effects on the variation. First, as previous research has found, longer noun phrase subjects disfavor contraction: the positive coefficient for host phrase length in words reflects increased likelihood of full form use with longer subjects. The tight credible intervals make this one of the most reliable predictors in the study.

Second, compared to human host phrases, non-human host phrases favor the use of the full form, replicating McLaughlin's (2014) analysis of a subset of the present data. Though collective host phrases also show a positive coefficient, the 95% credible intervals for this predictor cross 0: collective subjects aren't differentiated from human ones, but non-human subjects are.

Third, contra McLaughlin (2014), we also find an effect of host phrase proper nounhood. Full forms are disfavored (so, contraction is favored) with proper noun host phrases. This is a previously unobserved finding, but, when taken together with the effect of host phrase humanity, it is consistent with typological research, which has found that human and proper noun referents are at the higher end of an animacy scale together (Comrie, 1989, p. 195–196). These two predictors thus provide evidence that animate host phrases promote contraction. It's not immediately clear why an animacy effect should necessarily go in this direction, and most of the research on animacy effects in English concerns word order variation, rather than phonological reduction, so it cannot offer a useful precedent. What is clear, however, is that human and proper noun host phrases affect *is*-contraction in a comparable way, and their robust effects suggest that future work on *is*-contraction must include these predictors for a full account of the variation.

Fourth, all four levels of the preceding segment predictor show coefficients either below or above the 0 line. Preceding consonants favor full forms; preceding non-high vowels and /ɹ/ favor contracted forms. This corroborates Labov (1969), and suggests a pressure on *is*-contraction to maintain CVCV syllable patterning. It also suggests that post-vocalic /ɹ/ in American English behaves as a vowel, at least where contraction is concerned. The results also offer some support for Spencer's (2014) finding that contraction is sensitive to the voicing of a preceding consonant. The coefficient for voiceless consonants is higher than that for voiced consonants, although the credible intervals do overlap. The coefficient for a preceding high vowel, absent from **Figure 2** and **Table 3**, can be calculated to be −0.474 by summing the remaining coefficients and multiplying by −1; this gives us the following hierarchy of preceding segments on conditioning contraction, from most contraction to least:

> non-high vowel > /ɹ/ > high vowel > voiced consonant > voiceless consonant

The lower placement on the hierarchy of high vowels compared to non-high vowels corroborates my experience when coding that a preceding high vowel might lead the analyst to be more likely to hear a full form than otherwise. Still, its negative coefficient aligns with the other vocoids in favoring contracted over full forms.

The stress of syllables surrounding the contractable verb plays a minimal role, if any, in conditioning contraction. All

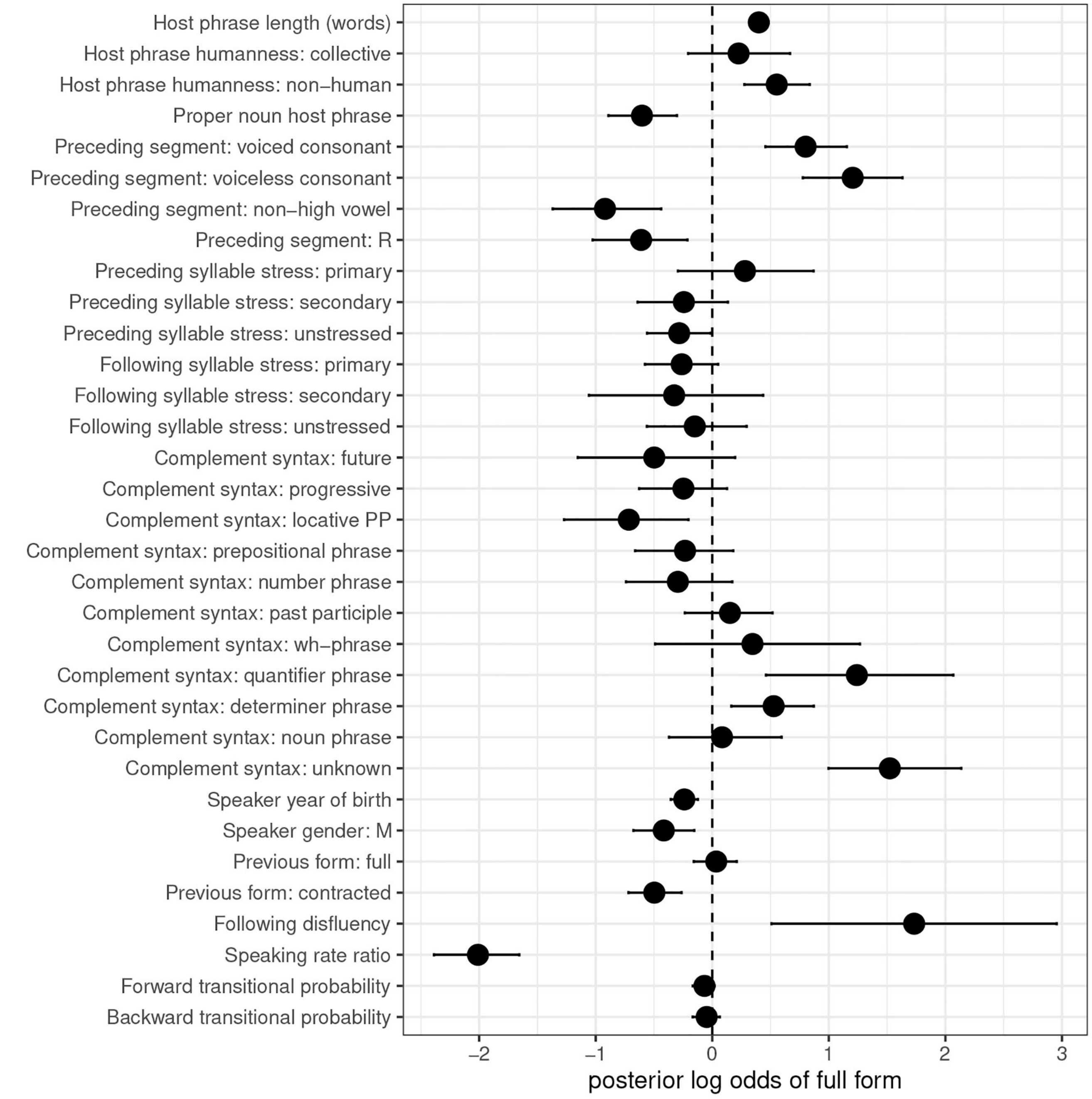


FIGURE 2 | Posterior means and 95% credible intervals for fixed effect predictors, *is*. Default level of dependent variable: contracted form. Points represent posterior log odds of the given predictor on use of the full form.

credible intervals for surrounding syllable stress levels cross 0, with the exception of unstressed preceding syllables, which display a negative coefficient that just avoids crossing the 0 line (upper bound: −0.014). This suggests a weak favoring effect of a preceding unstressed syllable on *is*-contraction, which is in keeping with Anttila's (2017) proposal that contraction will be more likely when the contractable verb is adjacent to an unstressed syllable. Anttila's (2017) proposal is also (weakly) supported by the positive coefficient of a preceding primary stressed syllable—favoring full over contracted forms—though the credible intervals cross 0 (with an 84% chance that the coefficient is positive). The effect of prosody on *is*-contraction deserves more careful consideration in future work. A first step would be to annotate surrounding syllable stress based on how syllables were actually phrased in production, rather than based on dictionary transcriptions of word stress in isolation.

TABLE 3 | Model estimates for predictors influencing *is*-contraction.

	Posterior mean	l-95% HPDI	u-95% HPDI	p(β < 0)	p(β > 0)
(Intercept)	1.6290118	0.8140859	2.3619368	0.000	1.000
Host phrase length (words)	0.3990877	0.3373437	0.4623129	0.000	1.000
Host phrase humanness: collective	0.2256141	−0.2071112	0.6676530	0.149	0.851
Host phrase humanness: non-human	0.5528309	0.2753459	0.8361011	0.000	1.000
Proper noun host phrase	−0.6033297	−0.8910730	−0.3031213	1.000	0.000
Preceding segment: voiced consonant	0.8008427	0.4555751	1.1543951	0.000	1.000
Preceding segment: voiceless consonant	1.2047807	0.7769507	1.6319018	0.000	1.000
Preceding segment: non-high vowel	−0.9206464	−1.3695070	−0.4383050	1.000	0.000
Preceding segment: R	−0.6110647	−1.0271768	−0.2129865	0.997	0.003
Preceding syllable stress: primary	0.2800213	-0.2950003	0.8693097	0.164	0.836
Preceding syllable stress: secondary	−0.2439706	−0.6405935	0.1341551	0.892	0.108
Preceding syllable stress: unstressed	−0.2840958	−0.5597557	−0.0136968	0.983	0.017
Following syllable stress: primary	−0.2628040	−0.5776638	0.0503263	0.948	0.052
Following syllable stress: secondary	−0.3271295	−1.0587807	0.4366904	0.798	0.202
Following syllable stress: unstressed	−0.1502357	−0.5612891	0.2949292	0.756	0.244
Complement syntax: future	−0.4980775	−1.1548223	0.1962291	0.915	0.085
Complement syntax: progressive	−0.2476477	−0.6272430	0.1268231	0.906	0.094
Complement syntax: locative PP	−0.7163825	−1.2717282	−0.2043430	0.995	0.005
Complement syntax: prepositional phrase	−0.2341470	−0.6619528	0.1804951	0.850	0.150
Complement syntax: number phrase	−0.2955229	−0.7405061	0.1714253	0.902	0.098
Complement syntax: past participle	0.1516862	−0.2355767	0.5163777	0.213	0.787
Complement syntax: wh-phrase	0.3449906	−0.4907484	1.2669180	0.226	0.774
Complement syntax: quantifier phrase	1.2402996	0.4602654	2.0678252	0.000	1.000
Complement syntax: determiner phrase	0.5265879	0.1631096	0.8713721	0.000	1.000
Complement syntax: noun phrase	0.0834676	−0.3707944	0.5944527	0.373	0.627
Complement syntax: unknown	1.5231752	0.9971601	2.1368757	0.000	1.000
Speaker year of birth	−0.2399602	−0.3561114	−0.1235707	1.000	0.000
Speaker gender: M	−0.4164843	−0.6765843	−0.1543645	0.998	0.002
Previous form: full	0.0339510	−0.1586210	0.2101696	0.379	0.621
Previous form: contracted	−0.4972574	−0.7191570	−0.2636629	1.000	0.000
Following disfluency	1.7319568	0.5081198	2.9550253	0.002	0.998
Speaking rate ratio	−2.0105721	−2.3881814	−1.6560303	1.000	0.000
Forward transitional probability	−0.0673087	−0.1676491	0.0205876	0.914	0.086
Backward transitional probability	−0.0482657	−0.1679330	0.0646831	0.785	0.215

Default form: contracted. Posterior means are log odds estimates of use of the full form. "l-95% HPDI" and "u-95% HPDI" are lower and upper bounds, respectively, of the 95% credible intervals, the areas in which 95% of the posterior probability density lies. "p(β < 0)" and "p(β > 0)" reflect the posterior probability that the coefficient of a given predictor is negative (favoring the contracted form) or positive (favoring the full form), respectively.

Complement syntax shows varying effects on *is*-contraction. Consistent with work on *is*-variation in AAVE, determiner phrases (which roughly map onto previous researchers' "noun phrase" category) favor full forms, while locative prepositional phrases favor contracted forms. Quantifier phrases, which are presumably also likely to have been called "noun phrases" in previous work (as they comprise complements such as *a little bit of everything, all these problems,* and *nothing*), pattern with determiner phrases in favoring full forms (though, surprisingly, noun phrases do not). At the same time, some classic contraction-favoring complements in previous work, such as future and progressive forms, show no reliable difference from 0, as do some new distinctions made for the present study, such as non-locative prepositional phrases, number phrases, past participles, and *wh*-phrases. It remains to be determined in future work whether collapsing any of these categories improves model fit. Anttila's (2017) proposal that the effect of complement syntax on *is*-contraction is an artifact of prosodic differences between complement types also deserves careful consideration. For the time being, one last observation worth noting is the strong positive coefficient of what were coded as "unknown" complements. Those cases where the speaker changed direction or restarted their sentence between uttering the verb and its complement strongly favor full forms. This finding has not previously been reported; I return to it later in this subsection.

The two sociodemographic predictors both have negative effects on the use of full forms. Rates of contraction are higher among younger speakers and among speakers identified by the corpus as male. This suggests an age-graded variable (Labov, 2001), though we know little about whether *is*-contraction carries social value, in the way that other age-graded variants are thought to do (Wagner, 2012). This remains an additional area for future research.

Contracted forms show persistence: a speaker who has just uttered a contracted form will be more likely to produce another one, relative to their likelihood of producing a contracted form as their first token of the conversation. There is no comparable persistence effect of full forms, however: production of a full form has no influence on whether a speaker will produce another. This is consistent with other work that has found the less frequent variant of a variable to trigger a stronger persistence effect than the more frequent one (see Tamminga, 2014, p. 97–117 for a review and some additional findings). Contracted forms surface only 36% of the time in the *is* data, suggesting that their persistence may be a surprisal effect.

A following disfluency (*uh* or *um*) strongly favors a speaker's using a full form. This finding is reminiscent of the full form-favoring effect of unknown complements mentioned three paragraphs earlier, and constitutes another previously-undocumented effect on contraction of *is*. One possible interpretation of these two findings is that a speaker's failure to plan the word following the target effectively causes the target to become phrase-final. Phrase-final position is an environment that disallows contraction (King, 1970). That said, there could also be a prosodic explanation: perhaps verbs in these environments bear more stress, disfavoring contraction (Anttila, 2017). The disfavoring effect of upcoming uncertainty on contraction should be probed further in future work; it connects to other research on the effects of production planning on sociolinguistic variation (e.g., Tanner et al., 2017).

This study is the first to find a clear, strong effect of speaking rate on *is*-contraction, with more contraction in faster speech relative to a speaker's average. This could be an effect of prosodic phrasing. There is some evidence that faster speech correlates with longer phrases (e.g., Quené, 2008). And if contraction requires the contractable verb and an adjacent word to be phrased together, as Anttila (2017) proposes, faster speech may make it more likely that speakers phrase their utterances in such a way that effects this. This is yet another indication that the effect of prosodic phrasing on contraction is a rich area for future study.

Finally, both of the measures of transitional probability show credible intervals that cross 0. This means that, in contrast to several previous studies, the Switchboard data show no predictability effects on *is*-contraction. However, a crucial distinction between those studies and this one is the stringent restriction on host phrases used here. I included no token whose host phrase was any sort of pronoun, with "pronoun" defined broadly to include indefinite pronouns. This contrasts, for instance, with Spencer (2014), who also restricted her data to non-pronominal subjects, but included indefinite pronouns among them—and found the expected predictability effects. This suggests that the predictability effects uncovered in previous work may in fact be better attributable to syntactic differences in the types of host phrases that were included in the data. A useful test would be to include tokens with indefinite pronoun hosts among the data used here, and see whether the transitional probability results change.

TABLE 4 | Confusion matrix for *has*-contraction.

	Predicted			
	Contracted	**Full**	**Intermediate**	**Total**
Observed				
Contracted	180	45	31	256
Full	52	173	36	261
Intermediate	52	58	72	182
Total	284	276	139	699

4.2. *Has*

4.2.1. Model Predictions and Accuracy

The goodness-of-fit statistics presented for the *is* model cannot be calculated for a multinomial model, but we can still examine the *has* model's predictive accuracy. **Table 4** shows the confusion matrix of predicted and observed forms for *has*. These predictions were obtained by identifying, for each data point, which of the three forms was most probable according to the model.

The rate of predictive accuracy for *has* is noticeably lower than it was for *is*, presumably a result of the smaller number of tokens and the difficulty imposed on the model of having to make three choices rather than two. The model predicts the correct form only 61% of the time, compared to the *is* model's 80%. Still, this 61% accuracy rate is a significant increase over simply guessing the most frequent form every time, which would give an accuracy rate of 37% ($p_{binom} < 0.001$).

Examining the recall rates shown in the rows of **Table 4**, we can see that the model does a much better job of predicting contracted forms (70% predicted correctly) and full forms (66% predicted correctly) than intermediate forms (only 40% predicted correctly). This may be attributable to the lower rate of representation of intermediate forms in the data (26% of observed forms compared to 37% for both full and contracted). But it also suggests that the predictors included in the present study, while reasonably appropriate for modeling occurrence of contracted and full forms of both verbs under study, may not be the best predictors for modeling occurrence of intermediate forms.

4.2.2. Results

The model of *is*-contraction allowed us to interrogate which factors condition the choice between contracted and full forms of *is*. By contrast, the multivariate model of *has*-contraction allows us to investigate the factors conditioning the choice between contracted and full forms, **and** the factors conditioning the choice between contracted and intermediate forms. However, under the analysis of *has*-variation proposed in MacKenzie (2013) and outlined in section 2.1, these two choices are, underlyingly, a single abstract choice. This means that, if that analysis is correct, the same factors should condition both choices: in other words, the two non-contracted forms of *has* should pattern together. Additionally, if contraction is conditioned in the same way regardless of verb, the same factors that favored *is*-contraction should be at play in *has*-contraction, and those same factors should favor contracted forms of *has* while disfavoring the other two.

Figure 3 provides a graphical representation of the estimated log odds of full form usage and intermediate form usage, with 95% credible intervals, for each predictor in the *has* model. Because there are two non-default forms to choose from, coefficients and credible intervals are presented for each. **Table 5** provides the posterior means, 95% Highest Posterior Density (credible) interval boundaries, and above-/below-0 probabilities for each predictor and each non-default variant. The top half of the table provides the coefficients associated with the choice between contracted and full for each predictor; the bottom half covers the choice between contracted and intermediate. The rows that say simply "Full" and "Intermediate" reflect intercept values: the log odds of using the indicated form over the contracted form when all predictors are set to their default level.

A first glance at **Figure 3** reveals wide credible intervals for nearly every point. Impressionistically, the credible intervals generally appear wider than those for *is*. This suggests more uncertainty in the *has* model, in keeping with its lower rate of predictive accuracy, and consistent with the smaller data set available for *has* compared to *is*.

Nonetheless, some clear effects are apparent. One of the most notable is the effect of host phrase length. Full and intermediate forms are both favored after longer host phrases, reflecting a disfavoring effect of long host phrases on contracted forms that matches that seen for *is*.

There is somewhat mixed evidence for the predictor of host phrase humanity. As was the case for *is*, both collective and non-human host phrases show positive coefficients on use of the non-default forms. As was also the case for *is*, the credible intervals for collective host phrases cross 0, for both non-default forms, suggesting no reliable difference in contraction rate between human and collective host phrases. But unlike what was the case for *is*, the credible intervals for the effect of non-human host phrases on full forms cross 0. Still, the model does output an 89% chance that the true coefficient for this predictor with this variant is greater than 0, i.e., positive. And the credible intervals for non-human host phrases on intermediate forms do not cross 0 (though they come very close to it, with a lower bound of 0.026). These findings suggest that host phrase animacy could be having the same effect on *has*-contraction as it has on *is*-contraction—that is, disfavoring contracted forms after non-human host phrases—but the results are inconclusive.

We find the same kind of result for proper noun host phrases. As with *is*, the coefficient for both non-default variants is negative. There is a 93% chance that proper noun host phrases favor the contracted variant over the full one, and a 78% chance that they favor the contracted variant over the intermediate one. But again, both credible intervals cross 0.

As with *is*, full forms of *has* are favored over contracted ones after consonants, particularly voiceless ones. (There is a 91% probability that voiced consonants favor full over contracted forms, though the credible interval crosses 0; there is a 98% probability that voiceless consonants favor full over contracted forms, though the credible interval approaches 0, with a lower bound of 0.042). In contrast to *is*, the disfavoring effect of vocoids (vowels and /ɹ/) on full forms of *has* is not in evidence—but preceding vowels do disfavor **intermediate** forms of *has*. All of these findings can be unified if we think of contraction as a phenomenon that seeks to minimize word-final consonant clusters and vowel-vowel hiatus. Consonant-final host words will disfavor contracted forms of any verb, which cliticize to their host word and create word-final consonant clusters. And vowel-final host words will disfavor vowel-initial verb forms which create hiatus: that is, full forms of *is*, and intermediate but not full forms of *has*. These ideas were first proposed by Labov (1969), and continue to find support in this larger, multi-verb data set.

As with *is*, surrounding syllable stress does not play a role in conditioning *has*-contraction. All credible intervals cross zero.

Where full forms are concerned, social factors behave in an identical way between the two verbs. Full forms of *has* are disfavored among younger speakers and by male speakers, as they were for *is*. But intermediate forms of *has* do not follow suit. The credible intervals for both predictors on intermediate forms cross 0, suggesting no influence of these predictors on use of intermediate forms, but a demonstrable influence on speakers' choice between contracted and full.

The persistence effect that can be demonstrated for *has*-contraction takes a different shape than that for *is*-contraction, where a previous contracted form boosted the likelihood of a speaker using a subsequent contracted form. Here, a previous full form boosts the likelihood of a speaker using a subsequent full form. I return to this discrepancy in section 5.

There is an 88% chance that *has* shows the same favoring of full forms in pre-disfluency position as *is*, though with only 14 pre-disfluency tokens in the *has* data, the model understandably shows uncertainty, with very wide confidence intervals. No comparable effect can be demonstrated for intermediate forms of *has*, but again, token counts are very low.

Finally, the last three predictors all accord with the results for *is*. A faster speaking rate relative to a speaker's baseline disfavors both full and intermediate forms, meaning that faster speech favors contracted forms—exactly the effect that was found for *is*. Neither measure of predictability has an effect on either of the non-default variants.

5. DISCUSSION

This paper started out with three questions:

1. Which factors condition the alternation between contracted and full forms of *is*?
2. Do those same factors condition the alternation between the contracted form of *has* and its two other possible forms?

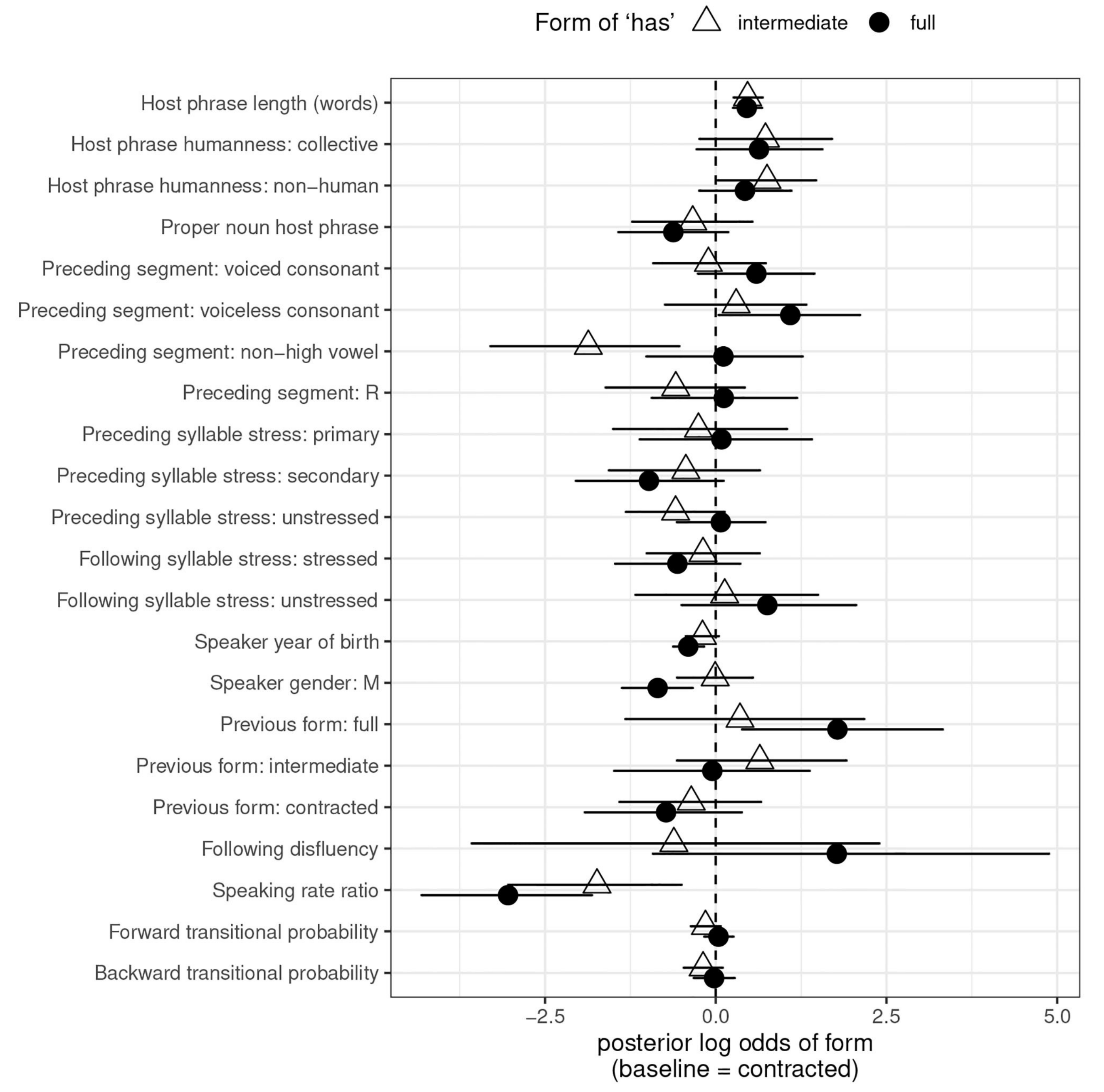


FIGURE 3 | Posterior means and 95% credible intervals for fixed effect predictors, *has*. Default level of dependent variable: contracted form. Points represent posterior log odds of the given predictor on use of the indicated form.

3. Does the patterning of *has* lend support to the analysis under which the three surface forms are derived from two stages of binary choice?

Speaking to questions 1 and 2, section 4 finds a number of predictors, some novel and some well-documented, to condition variation in *is*. These include phonological and semantic properties of the verb's host phrase, speaker sociodemographic factors, and characteristics of the speaking situation, such as speech rate and persistence. Most of these also affect variation in *has*, in similar ways. Specifically, of the thirteen predictors examined for both verbs, six of them have non-null effects on both (host phrase length, preceding segment, year of birth, sex/gender, persistence, and speaking rate); a further four have null effects on both (preceding syllable stress, following syllable stress, forward transitional probability, backward transitional probability); and the remaining three show the same patterning

TABLE 5 | Model estimates for predictors influencing *has*-contraction.

	Posterior mean	l-95% HPDI	u-95% HPDI	p(β < 0)	p(β > 0)
Full	1.7562520	-0.4610893	4.1016186	0.0618182	0.9381818
Full - Host phrase length (words)	0.4540251	0.2547957	0.6774660	0.0000000	1.0000000
Full - Host phrase humanness: collective	0.6331986	−0.2787359	1.5609129	0.0840909	0.9159091
Full - Host phrase humanness: non-human	0.4264335	−0.2411717	1.1071036	0.1100000	0.8900000
Full - Proper noun host phrase	−0.6217921	−1.4240903	0.1825051	0.9345455	0.0654545
Full - Preceding segment: voiced consonant	0.5942323	−0.2611305	1.4441186	0.0922727	0.9077273
Full - Preceding segment: voiceless consonant	1.0910646	0.0423456	2.1108090	0.0200000	0.9800000
Full - Preceding segment: non-high vowel	0.1120166	−1.0148022	1.2719679	0.4218182	0.5781818
Full - Preceding segment: R	0.1172536	−0.9374153	1.1907104	0.4131818	0.5868182
Full - Preceding syllable stress: primary	0.0828608	−1.1125612	1.4071405	0.4618182	0.5381818
Full - Preceding syllable stress: secondary	−0.9788212	−2.0480402	0.1140841	0.9627273	0.0372727
Full - Preceding syllable stress: unstressed	0.0735986	−0.5686875	0.7261148	0.4154545	0.5845455
Full - Following syllable stress: stressed	−0.5622588	−1.4733034	0.3584011	0.8863636	0.1136364
Full - Following syllable stress: unstressed	0.7545892	−0.4991466	2.0553168	0.1136364	0.8863636
Full - Speaker year of birth	−0.4034372	−0.6235081	−0.1729058	1.0000000	0.0000000
Full - Speaker gender: M	−0.8515252	−1.3734524	−0.3390654	0.9986364	0.0013636
Full - Previous form: full	1.7816588	0.3862631	3.3237094	0.0072727	0.9927273
Full - Previous form: intermediate	−0.0513835	−1.4860966	1.3737515	0.5186364	0.4813636
Full - Previous form: contracted	−0.7282325	−1.9161779	0.3802393	0.8909091	0.1090909
Full - Following disfluency	1.7721448	−0.9199031	4.8789925	0.1163636	0.8836364
Full - Speaking rate ratio	−3.0423378	−4.3067556	−1.8131940	1.0000000	0.0000000
Full - Forward transitional probability	0.0393089	−0.1678315	0.2581302	0.3640909	0.6359091
Full - Backward transitional probability	−0.0251222	−0.3218068	0.2756846	0.5700000	0.4300000
Intermediate	−0.1687226	−2.4067836	2.2045201	0.5613636	0.4386364
Intermediate - Host phrase length (words)	0.4653107	0.2632404	0.6853951	0.0000000	1.0000000
Intermediate - Host phrase humanness: collective	0.7277026	−0.2379348	1.7027128	0.0686364	0.9313636
Intermediate - Host phrase humanness: non-human	0.7504033	0.0264793	1.4700295	0.0218182	0.9781818
Intermediate - Proper noun host phrase	−0.3374735	−1.2233850	0.5361169	0.7795455	0.2204545
Intermediate - Preceding segment: voiced consonant	−0.1081011	−0.9152518	0.7319900	0.5981818	0.4018182
Intermediate - Preceding segment: voiceless consonant	0.2983763	−0.7416489	1.3259396	0.2890909	0.7109091
Intermediate - Preceding segment: non-high vowel	−1.8667305	−3.2993466	−0.5337918	0.9968182	0.0031818
Intermediate - Preceding segment: R	−0.5862945	−1.6123424	0.4234284	0.8695455	0.1304545
Intermediate - Preceding syllable stress: primary	−0.2542370	−1.5032514	1.0433443	0.6413636	0.3586364
Intermediate - Preceding syllable stress: secondary	−0.4382973	-1.5653899	0.6473658	0.7918182	0.2081818
Intermediate - Preceding syllable stress: unstressed	−0.5886084	−1.3148520	0.1284327	0.9450000	0.0550000
Intermediate - Following syllable stress: stressed	−0.1869565	−1.0106447	0.6422665	0.6736364	0.3263636
Intermediate - Following syllable stress: unstressed	0.1303545	−1.1750509	1.4957722	0.4209091	0.5790909
Intermediate - Speaker year of birth	−0.1942786	−0.4365133	0.0420955	0.9463636	0.0536364
Intermediate - Speaker gender: M	−0.0088502	−0.5666323	0.5440431	0.5104545	0.4895455
Intermediate - Previous form: full	0.3546801	−1.3224632	2.1734082	0.3563636	0.6436364
Intermediate - Previous form: intermediate	0.6451244	−0.5674888	1.9129641	0.1536364	0.8463636
Intermediate - Previous form: contracted	−0.3588215	−1.4097452	0.6613422	0.7500000	0.2500000
Intermediate - Following disfluency	−0.6136116	−3.5729961	2.3940366	0.6559091	0.3440909
Intermediate - Speaking rate ratio	−1.7394001	−3.0362574	−0.5022914	0.9968182	0.0031818
Intermediate - Forward transitional probability	−0.1509607	−0.3625895	0.0697880	0.9172727	0.0827273
Intermediate - Backward transitional probability	−0.1857577	−0.4676284	0.1008418	0.8981818	0.1018182

Default form: contracted. Posterior means are log odds estimates of use of the indicated form. "l-95% HPDI" and "u-95% HPDI" are lower and upper bounds, respectively, of the 95% credible intervals, the areas in which 95% of the posterior probability density lies. "p(β < 0)" and "p(β > 0)" reflect the posterior probability that the coefficient of a given predictor is negative (favoring the contracted form) or positive (favoring the indicated form), respectively. "Full" and "Intermediate" reflect intercept values.

for both verbs, but the *has* credible intervals cross the 0 line (host phrase humanness, host phrase proper nounhood, following disfluency).

But the answer to the third question is not as straightforward. To recap, MacKenzie (2013) provided an analysis of intermediate forms of *has* under which they were derived from full forms of *has*. Under this analysis, three forms were derived via two binary choices: first, a choice between contracted and full; second, a choice between full and intermediate. This is a common approach in variationist sociolinguistics to modeling three-way variation (Sankoff and Rousseau, 1989), and MacKenzie (2013) drew on two pieces of evidence to support it in the case of contraction. First, contracted forms are used at near-identical rates for *is* and for *has*. This is consistent with there being a first stage of choice between contracted and any other forms, and with this first stage of choice having an identical rate of application across verbs. As was shown at the beginning of section 4, this holds up in the present data. The second piece of evidence was the patterning of forms by host phrase length. Contracted forms of both verbs were disfavored after longer host phrases, while both full forms and intermediate forms (of *has*) were favored. Again, this holds up in the present data.

Additional support for this analysis in the present study comes from speaking rate, another factor that patterns in the same way: faster speech favors contracted forms of *is* over full forms, and contracted forms of *has* over full and intermediate forms. Weaker evidence in support of the two-stage analysis comes from the predictors of host phrase humanness, host phrase proper nounhood, and year of birth. For all of these, the model coefficients for full and intermediate forms of *has* have the same polarity as each other and as the full form of *is* (either all positive or all negative), but the credible intervals for *has* cross 0 for most of them.

But not all predictors pattern with a contracted vs. full + intermediate split, as the two-stage analysis would predict. In fact, each possible grouping of the three variants of *has* is attested in the data. To discern this, I ran a second multinomial model on the *has* data, with intermediate forms set as the default (reference) level. In this model, coefficients for full forms tell us which factors condition the choice between full and intermediate—the second-stage choice under MacKenzie's (2013) model. The results of this second *has* model are omitted for space reasons, but are available in the **Supplementary Material.**

This second model reveals that, for some predictors, contracted and intermediate forms pattern together in opposition to full forms. This holds for speaker gender: male speakers favor contracted forms over full **and** intermediate forms over full. This suggests that the gender effect on contraction operates on a distinction between full forms and forms that are phonologically reduced in some way. Even more interesting is the attested persistence effect for *has*. Recall that the persistence effect for *is* showed contracted forms begetting contracted forms, but without a concomitant persistence effect for full forms, which I attributed to contracted forms' being the less commonly used variant of *is*. The persistence effect for *has* is as follows: when a previous form is full, full forms are more likely compared to both contracted and intermediate forms, but neither of the other two variants triggers any persistence effects itself. As with the gender effect, this is interpretable as a full vs. reduced split in variant patterning. And, analogous to what we found for *is*, full forms are in the minority when we split the variants in that way: 37% of forms of *has* are full, compared to 63% which are reduced (i.e., contracted or intermediate). We can unite both verbs' persistence behavior by saying that persistence operates on a full vs. reduced division of variants, and takes the shape of the minority variant in this dichotomy triggering further instances of itself.

Finally, one predictor operates on intermediate as opposed to contracted and full forms. This is preceding segment, specifically, the effect of a preceding vowel. Preceding vowels favor contracted and full forms over intermediate forms of *has*, but play no role in the choice between contracted and full. This again makes sense in light of what was found for *is*, where a preceding vowel disfavored the use of full forms, likely due to a hiatus-avoidance strategy. For *has*, the intermediate form—the only one of the three that is vowel-initial—is disfavored after vowels, but a preceding vowel has no effect on whether a speaker will choose either of the two consonant-initial forms.

These findings complicate the original analysis of *has*-variation put forth by MacKenzie (2013). On the one hand, some predictors do support the proposal that speakers first make a choice between a contracted form of *has* and a full form, which may or may not become intermediate at a later stage of the derivation. On the other hand, effects like the one for a preceding vowel cannot be accommodated. This predictor shapes the choice between contracted and intermediate forms, but under the two-stage analysis, there is no point at which a speaker ever has to decide between using a contracted or an intermediate form. Intermediate forms haven't been derived at the point at which a speaker chooses whether to use a contracted form or not. And yet the findings show us that certain conditioning factors do operate on such decisions.

All of this appears to suggest that variation in *has* is a three-way choice for speakers, between full, intermediate, and contracted forms. But different predictors favor or disfavor different types of forms. Some predictors operate on the distinction between full and phonologically reduced forms: that is, between full on the one hand, and intermediate and contracted on the other. Other predictors are sensitive to whether a form is vowel-initial or not, operating on the distinction between intermediate on the one hand, and full and contracted on the other. And a final set of environments—those that were originally taken to support the two-stage analysis, because they show contracted forms patterning in opposition to full and intermediate ones—can be interpreted as operating on the distinction between non-syllabic and syllabic forms. This last set of environments is perhaps the most interesting one, because the apparent syllabicity effect suggests a prosodic aspect to the variation. And, indeed, the predictors that are sensitive to variant syllabicity include host phrase length and speaking rate, both of which may have their source in prosodic phrasing (Quené, 2008; Anttila, 2017).

As a result, the *has*-contraction findings cast the *is*-contraction findings in a new light. Studying *has* introduces a third form, the intermediate form, which is syllabic (like full), phonologically reduced (like contracted), and vowel-initial (like neither). Observing how it patterns with respect to the other two variants can help us understand which attributes of a form the conditioning factors are sensitive to. For instance, by studying persistence effects on the two verbs, we learn that persistence appears to operate over a phonologically full vs. phonologically reduced dichotomy, with whichever form is in the minority of these two categories triggering subsequent instances of itself. Without the data from *has*, the persistence effects on *is* would be ambiguous between this interpretation and two other interpretations: one in which persistence operates on a vowel-initial vs. consonant-initial dichotomy, and one in which persistence operates on a syllabic vs. non-syllabic dichotomy. Comparing *has*-contraction to *is*-contraction has thus given us deeper insight into how the mechanisms that constrain contraction operate.

6. CONCLUSION

This paper has examined variation in phonological form of two tensed verbs in English, *is* and *has*. Both verbs variably surface in a single-consonant contracted form and a form with all phonological material intact. *Has* differs from *is* in allowing a third form, which is reduced compared to the full form of *has*, making it phonologically near-identical to the full form of *is*. This raises questions about whether the different forms of the two verbs will pattern similarly to one another with respect to a number of internal and external factors. And this, in turn, can inform our analysis of how these different forms are related to one another.

I find a number of similarities in the patterning of the two verbs. These include the overall rate at which the contracted form is used, the constraints that affect the variation, and which form(s) those constraints favor. For both of these verbs, this study has uncovered un(der)documented effects on contraction which deserve further investigation, such as the favoring effect of host phrase animacy on contracted forms, and the potential effect of prosodic phrasing in shaping speakers' choice between syllabic and non-syllabic variants. This latter finding connects with other recent work urging more consideration of the role of prosodic information in conditioning variable processes (Kendall, 2013; Tanner et al., 2017).

But I hope the most lasting contribution of this work will be a methodological one. I approached the ternary variation shown by *has* not by grouping the variants into a particular binary opposition, but by allowing them to vary independently in a Bayesian MCMCglmm. And indeed, by doing this, I uncovered evidence that all three logically possible binary oppositions are evident in the data to some degree. This cannot be captured by modeling *has*-contraction as two binary choices, but rather suggests a three-way choice. Ternary variation like this raises important questions about the nature of the linguistic variable, and complicates the "single-input, single-output" formula so common in traditional variationist sociolinguistic research. It is my hope that more researchers working with non-binary variables will make use of the methods employed here, allowing potential variant groupings to come out of the data rather than imposing groupings on the data themselves.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was supported by faculty research funds from New York University.

ACKNOWLEDGMENTS

Thanks to Kimberley Baxter, Guy Tabachnick, and members of the Phonetics and Experimental Phonology and Sociolinguistics Labs at New York University. Jason Grafmiller generously made available code that was used for analyses, figures, and tables.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00058/full#supplementary-material

REFERENCES

Anttila, A. (2016). Phonological effects on syntactic variation. *Annu. Rev. Linguist.* 2, 115–137. doi: 10.1146/annurev-linguistics-011415-040845

Anttila, A. (2017). "Stress, phrasing, and auxiliary contraction in English," in *The Morphosyntax-Phonology Connection: Locality and Directionality at the Interface*, eds V. Gribanova and S. S. Shih (New York, NY: Oxford University Press), 143–170. doi: 10.1093/acprof:oso/9780190210304.003.0006

Barth, D., and Kapatsinski, V. (2014). A multimodel inference approach to categorical variant choice: Construction, priming and frequency effects on the choice between full and contracted forms of *am*, *are* and *is*. *Corpus Linguist. Linguist. Theory* 13, 203–260. doi: 10.1515/cllt-2014-0022

Bresnan, J. (2018). *Formal grammar, usage probabilities, and English tensed auxiliary contraction* (MS thesis). Stanford University, Stanford, CA, United States.

Cedergren, H., and Sankoff, D. (1974). Variable rules: performance as a statistical reflection of competence. *Language* 50, 333–355. doi: 10.2307/412441

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. Chicago, IL: University of Chicago Press.

Dilley, L., Gamache, J., Wang, Y., Houston, D. M., and Bergeson, T. R. (2019). Statistical distributions of consonant variants in infant-directed speech: evidence that /t/ may be exceptional. *J. Phonet*. 75, 73–87. doi: 10.1016/j.wocn.2019.05.004

Drummond, R. (2011). Glottal variation in /t/ in non-native English speech: patterns of acquisition. *English World Wide* 32, 280–308. doi: 10.1075/eww.32.3.02dru

Finegan, E., and Biber, D. (1986). "Two dimensions of linguistic complexity in English," in *Social and Cognitive Perspectives on Language: Southern California Occasional Papers in Linguistics 11*, eds J. Connor-Linton, C. J. Hall, and M. McGinnis (Los Angeles, CA: Dept. of Linguistics, University of Southern California), 1–24.

Foulkes, P., Docherty, G., and Watt, D. (2005). Phonological variation in child-directed speech. *Language* 81, 177–206. doi: 10.1353/lan.2005.0018

Frank, A., and Jaeger, T. F. (2008). "Speaking rationally: uniform information density as an optimal strategy for language production," in *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)* (Washington, DC: Frank & Jaeger), 939–944.

Godfrey, J. J., and Holliman, E. (1997). *Switchboard-1 Release 2*. Philadelphia, PA: Linguistic Data Consortium.

Grafmiller, J., Szmrecsanyi, B., and Hinrichs, L. (2018). Restricting the restrictive relativizer: constraints on subject and non-subject English relative clauses. *Corpus Linguist. Linguist. Theory* 14, 309–355. doi: 10.1515/cllt-2016-0015

Haddican, W., and Zweig, E. (2012). The syntax of manner quotative constructions in English and Dutch. *Linguist. Variat*. 12, 1–26. doi: 10.1075/lv.12.1.01had

Hadfield, J. D. (2010a). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw*. 33, 1–22. doi: 10.18637/jss.v033.i02

Hadfield, J. D. (2010b). *MCMCglmm: Markov chain Monte Carlo methods for generalised linear mixed models* (MS thesis). University of Edinburgh, Edinburgh, United Kingdom.

Hadfield, J. D. (2019). *MCMCglmm course notes* (MS thesis). University of Edinburgh, Edinburgh, United Kingdom.

Hughes, A., Trudgill, P., and Watt, D. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, 5th Edn*. London: Hodder Education.

Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Lang. Linguist. Compass* 3, 359–383. doi: 10.1111/j.1749-818X.2008.00108.x

Kendall, T. (2013). *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. New York, NY: Springer. doi: 10.1057/9781137291448

King, H. V. (1970). On blocking the rules for contraction in English. *Linguist. Inq*. 1, 134–136.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762. doi: 10.2307/412333

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Malden, MA: Blackwell.

Levshina, N. (2015a). *Bayesian Logistic Models With MCMCglmm: A Brief Tutorial*. Leuven University.

Levshina, N. (2015b). *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Philadelphia, PA: John Benjamins Publishing Company. doi: 10.1075/z.195

Levshina, N. (2016). When variables align: a Bayesian multinomial mixed-effects model of English permissive constructions. *Cogn. Linguist*. 27, 235–268. doi: 10.1515/cog-2015-0054

MacKenzie, L. (2012). *Locating variation above the phonology* (Ph.D. thesis). Philadelphia, PA: University of Pennsylvania.

MacKenzie, L. (2013). Variation in English auxiliary realization: a new take on contraction. *Lang. Variat. Change* 25, 17–41. doi: 10.1017/S0954394512000257

MacKenzie, L. (2016). "Production planning effects on variable contraction in English," in *University of Pennsylvania Working Papers in Linguistics 22.2: Selected Papers from NWAV 44*, ed H. Jeoung (Philadelphia, PA: MacKenzie), 121–130.

McElhinny, B. S. (1993). Copula and auxiliary contraction in the speech of White Americans. *Am. Speech* 68, 371–399. doi: 10.2307/455773

McLaughlin, B. (2014). *Animacy in morphosyntactic variation* (Ph.D. thesis). University of Pennsylvania, Philadelphia, PA, United States.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *J. Acous. Soc. Am*. 123, 1104–1113. doi: 10.1121/1.2821762

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rickford, J. R., Ball, A., Blake, R., Jackson, R., and Martin, N. (1991). Rappin on the copula coffin: theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Lang. Variat. Change* 3, 103–132. doi: 10.1017/S0954394500000466

Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language* 81, 613–644. doi: 10.1353/lan.2005.0149

Sankoff, D., and Rousseau, P. (1989). Statistical evidence for rule ordering. *Lang. Variat. Change* 1, 1–18. doi: 10.1017/S0954394500000090

Sharma, D., and Rickford, J. R. (2009). AAVE/Creole copula absence: a critique of the imperfect learning hypothesis. *J. Pidgin Creole Lang*. 24, 53–90. doi: 10.1075/jpcl.24.1.03sha

Spencer, J. D. (2014). *Stochastic effects in the grammar: toward a usage-based model of copula contraction* (Ph.D. thesis). Stanford University, Stanford, CA, United States.

Straw, M., and Patrick, P. L. (2007). Dialect acquisition of glottal variation in /t/: Barbadians in Ipswich. *Lang. Sci*. 29, 385–407. doi: 10.1016/j.langsci.2006.12.025

Szmrecsanyi, B., Biber, D., Egbert, J., and Franco, K. (2016). Toward more accountability: modeling ternary genitive variation in Late Modern English. *Lang. Variat. Change* 28, 1–29. doi: 10.1017/S0954394515000198

Tagliamonte, S. A. (2006). *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801624

Tamminga, M. (2014). *Persistence in the production of linguistic variation* (Ph.D. thesis). University of Pennsylvania, Philadelphia, PA, United States.

Tamminga, M. (2016). Persistence in phonological and morphological variation. *Lang. Variat. Change* 28, 335–356. doi: 10.1017/S0954394516000119

Tanner, J., Sonderegger, M., and Wagner, M. (2017). Production planning and coronal stop deletion in spontaneous speech. *Lab. Phonol. J. Assoc. Lab. Phonol*. 8, 1–39. doi: 10.5334/labphon.96

Wagner, S. E. (2012). Age grading in sociolinguistic theory. *Lang. Linguist. Compass* 6, 371–382. doi: 10.1002/lnc3.343

Weide, R. (2008). *The CMU Pronouncing Dictionary*. Pittsburgh, PA: Weide. Carnegie Mellon University.

Wolfram, W. A. (1991). The linguistic variable: fact and fantasy. *Am. Speech* 66, 22–32. doi: 10.2307/455432

Wolk, C., Bresnan, J., Rosenbach, A., and Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* 30, 382–419. doi: 10.1075/dia.30.3.04wol

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

One Hundred Years of Migration Discourse in *The Times*: A Discourse-Historical Word Vector Space Approach to the Construction of Meaning

Lorella Viola[1]* **and Jaap Verheul**[2]

[1] Luxembourg Centre for Contemporary and Digital History (C^{2}DH), University of Luxembourg, Esch-sur-Alzette, Luxembourg, [2] Department of History and Art History, Utrecht University, Utrecht, Netherlands

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Heike Wiese,
Humboldt University of Berlin, Germany
Chaya Liebeskind,
Jerusalem College of Technology, Israel

****Correspondence:***
Lorella Viola
lorella.viola@uni.lu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *05 March 2020*
Accepted: *17 July 2020*
Published: *10 September 2020*

Citation:
Viola L and Verheul J (2020) One Hundred Years of Migration Discourse in The Times: A Discourse-Historical Word Vector Space Approach to the Construction of Meaning. Front. Artif. Intell. 3:64. doi: 10.3389/frai.2020.00064

This study proposes an experimental method to trace the historical evolution of media discourse as a means to investigate the construction of collective meaning. Based on distributional semantics theory (Harris, 1954; Firth, 1957) and critical discourse theory (Wodak and Fairclough, 1997), it explores the value of merging two techniques widely employed to investigate language and meaning in two separate fields: neural word embeddings (computational linguistics) and the discourse-historical approach (DHA; Reisigl and Wodak, 2001) (applied linguistics). As a use case, we investigate the historical changes in the semantic space of public discourse of migration in the United Kingdom, and we use the *Times Digital Archive* (TDA) from 1900 to 2000 as dataset. For the computational part, we use the publicly available TDA word2vec models[1] (Kenter et al., 2015; Martinez-Ortiz et al., 2016); these models have been trained according to sliding time windows with the specific intention to map conceptual change. We then use DHA to triangulate the results generated by the word vector models with social and historical data to identify plausible explanations for the changes in the public debate. By bringing the focus of the analysis to the level of discourse, with this method, we aim to go beyond mapping different senses expressed by single words and to add the currently missing sociohistorical and sociolinguistic depth to the computational results. The study rests on the foundation that social changes will be reflected in changes in public discourse (Couldry, 2008). Although correlation does not prove direct causation, we argue that historical events, language, and meaning should be considered as a mutually reinforcing cycle in which the language used to describe events shapes explicit meanings, which in turn trigger other events, which again will be reflected in the public discourse.

Keywords: word-vector space, migration discourse, historical newspapers, critical discourse analysis, diachronic conceptual change, language & media, migration history

[1] https://doi.org/10.5281/zenodo.1494140

INTRODUCTION

The emergence of unprecedented masses of digital data has brought an upsurge in Natural Language Processing (NLP) studies concerned with language and meaning. These studies today are mostly based on distributional semantics theory (Harris, 1954; Firth, 1957) and typically use techniques such as neural word embeddings to map different senses expressed by single words. However, some computational linguists have observed that, with the exception of few recent initiatives that go beyond single words[2], most current methods have failed to adopt more holistic approaches. A recent survey of studies on lexical semantic change detection (i.e., Tahmasebi et al., 2018), for instance, has indicated that the "issue of interdependence between semantic changes of different words" remains largely unexplored (Tahmasebi et al., 2018, p. 42). This would be due to that fact that works on lexical semantic change based on neural word embeddings have almost exclusively investigated single words. According to these studies, meaning change should on the contrary be understood as belonging to "an intricate net of word-to-word interrelation" as the focus on single words does not allow for a comprehensive view of how a given word changes meaning. This may suggest that, rather than looking at word senses separately, whole concepts or topics should be the focus of inquiry so that meaning changes are studied in the context of other words that express (or used to express) the same or related concepts.

Another exciting challenge that the field of NLP on language and meaning still presents concerns the scope of the inquiry. Kutuzov et al. (2018), for instance, have pointed out how the investigation continues to be concerned more with proving that a change has happened, rather than with identifying potential explanations for it. These authors, for instance, noted that "a more detailed analysis of the nature of the shift is needed," and similar to Tahmasebi et al. (2018), this could be accomplished through the identification of "groups of words that shift together in correlated ways," or with "identifying the source of a shift" (Tahmasebi et al., 2018, p. 1393), for instance, by studying linguistic or extralinguistic causes.

The availability of unprecedented masses of digital data has also brought the issue of interdisciplinarity between the humanities and the sciences at the center of the academic debate. It is certainly true that over the past few years, we have witnessed a growing number of studies that have combined approaches and methods from both fields. In disciplines such as linguistics and history, for example, the digital turn has called for a reconceptualization of the practice, almost forcing scholars to adopt advanced quantitative methods in their research, whereas in the humanities at large, it has led to the emergence of completely new fields such as digital humanities. Conversely, in computational linguistics, scholars have increasingly integrated linguistic theories and social data into their models, allowing for an increase in performance as well as significant advances, even in neighbor fields such as machine learning. Despite the major achievements, however, scholars across both sides have expressed a need for more integrated methods to combine perspectives, as well as accelerate and expand knowledge, as they feel that currently expertise remains essentially disconnected (e.g., Jockers, 2013; Snow, 2013; Tahmasebi et al., 2018). This struggle to actualize interdisciplinarity, particularly in the humanities, is perhaps best reflected in the difficulty to combine close reading with distant reading (Viola and Verheul, 2019). While computer science delivers solutions to automate or semiautomate analytical processes, close-reading approaches continue to be largely preferred in linguistics and the humanities. At the same time, scholars, particularly those working in digital humanities, have attempted to blend both approaches upon the conviction that "quantitative methods are most effective when used alongside the close textual reading" (Gooding et al., 2012, 2013). Jockers (2013), for instance, suggests what he calls a "macroanalysis" approach, whereas Graham et al. (2016) propose a workbench of different tools, called the "historians' macroscope." Similarly, Lee (2019) uses a range of distant reading techniques to automate a large part of the data preprocessing destined to critical discourse analysis (CDA) investigations, and finally, Viola and Verheul (2019) argue for a merged method, "discourse-driven topic modeling," effective at uncovering and making sense of historical patterns in large quantities of textual data. What these scholars have essentially tried to achieve is a mutual compensation for the limitations of both distant and close reading approaches: the need for an analytical contextualization of quantitative findings of the former and the impossibility of critically reading everything of the latter (Viola and Verheul, 2019).

This study aims to address such current challenges by exploring the value of merging two different techniques widely employed in two separate fields: neural word embeddings, a quantitative, distant reading method used in computational linguistics to investigate meaning change; and the discourse-historical approach (DHA; Reisigl and Wodak, 2001), a qualitative, close reading methodology used in applied linguistics to investigate the relationship between language and discourse. With this experimental method, our goal is to widen the focus of the analysis from identifying that a word has changed its meaning over time to exploring plausible extralinguistic factors that may reveal the mechanisms involved in the construction of collective meaning. As a use case, we investigate the historical changes in the semantic space of public discourse of migration in the United Kingdom, and we use the *Times Digital Archive* (TDA) from 1900 to 2000 as dataset. For the computational part, we use the publicly available TDA word2vec models (Kenter et al., 2015; Martinez-Ortiz et al., 2016); these models have been trained according to sliding time windows (cfr. *Methodology and Dataset*) with the specific intention to map conceptual change. We then use DHA to triangulate the results generated

[2] A recent attempt to overcome such limitation has been proposed by *deep contextualized word embeddings* such as ELMo (Embeddings from Language Models; Peters et al., 2018), which go beyond traditional embedding techniques. The innovation consists in considering an entire sentence—as opposed to single words—before assigning an embedding to each constituting word. While this approach certainly marks an important step forward in NLP and appears to improve the performance of automated processes such as word-sense disambiguation, sentiment analysis, etc., it may not be the best approach to map the changing meaning of words over time as the problem of historicizing the meaning of the words in context remains.

by the word vector models with social and historical data in order to provide plausible clarifications of the processes underpinning the construction of meaning itself. By merging these two complementary techniques, we hope to bring together different research modalities that could provide scholars with a research tool informed by critical, methodological, and empirical approaches (Berry and Fagerjord, 2017, p. 104).

The proposed method should not be seen as necessarily "better" than others; instead, our intention is to propose a more holistic approach that goes beyond simply identifying meaning changes and that could provide avenues for understanding the mechanisms underlying the construction of meaning in relation to language and public discourse. By integrating the quantitative findings with a discourse-historical interpretation, the method intends to add the currently missing sociohistorical and sociolinguistic depth to the computational results. This combination allows us to see how public discourse around the topic of migration has changed over the course of one century and how such changes may reflect the underlying sociohistorical events. We argue that this kaleidoscopic approach may reveal more than just the occurrence of a change in meaning: the method may help us to discover why such change has occurred at a given time and, more widely, how meaning is collectively constructed.

LANGUAGE AND THE CONSTRUCTION OF MEANING

The study of the relationship between language and meaning has long been the interest of many disciplines, including philosophy, psychology, anthropology, history, literature, linguistics, and, more recently, computational linguistics. As a result, authors have proposed a wide range of terminologies, classifications, and definitions. Although these publications have been influential in deepening our understanding of the complexities of word meaning, they also produced a multiplicity of labels and taxonomies that sometimes caused disagreement. For instance, there is no consensus on the outstanding question whether we should theoretically distinguish between *meaning* and *concept* or whether we should use these terms interchangeably. In fields such as conceptual history, for example, *concepts* seem to have been considered simply as words (Tahmasebi et al., 2018, p. 3). This would be reflected in the argument that the change in the vocabulary of specific terms indicates a change in the way the respective societal groups use such terms (e.g., Brunner et al., 1972; Koselleck, 1992, 2004); hence, in this field, definitions of *concept* are typically fuzzy and contested (Margolis and Laurence, 2005; Tahmasebi et al., 2018). This nevertheless important work conducted by conceptual historians on conceptual change (e.g., Skinner, 1969, 1978, 2012; De Bolla, 2013; Gavin, 2015; Recchia et al., 2017; De Bolla et al., 2019) has primarily searched for evidence of variation (i.e., vocabulary change), which is looked at through the lens of historical events or long-term changes in social stratification of society (Koselleck, 2004). Therefore, fundamental alterations in the meaning of keywords are interpreted as the reflection of conceptual changes (Pocock, 2016). We aim to expand on such studies by widening the scope of the investigation and focusing on what changes in society and language tell us about how meaning itself is constructed, which has often fallen beyond the scope of historical inquiry. We argue that this approach may build an interdisciplinary bridge between linguistics and conceptual history in studying the relation between changes in language and those in society.

Historical linguists, too, have identified culture as a crucial factor in language change; at the same time, however, they also point at even more fundamental mechanisms that trigger language change, such as for example, language contact. Linguists believe that before looking at potential changes in the meaning of words and what such changes might reveal, it is first essential to ask what meaning is in relation to language and society. In this sense, the most substantial theoretical and empirical contribution is offered by linguistics and, more recently, by modern empirical linguistics (i.e., corpus linguistics and computational linguistics). For this reason, we will here only refer to work within these fields that has attempted to address these questions.

Linguistic Approaches to the Study of Language and Meaning

Before the advent of powerful computers and the availability of large historical linguistic datasets, language was mostly studied through invented examples, using a speculative, intuitive approach. With the exception of few sporadic pioneering initiatives, for instance, in dialectology (i.e., Wenker, 1878) or in the first wave of variation studies (e.g., Labov, 1966; Trudgill, 1974; Macaulay, 1977), linguists primarily used introspective language competence and perception to formulate theories of what was possible and not possible in a language. Typically, they would formulate explanatory theories to describe certain phenomena and then invent examples that would confirm those theories. Because gaining access to real language data was costly and very time-consuming, their knowledge of the language as native speakers was the preferred "data" they would use. Consequently, these imagined examples of what was *possible* in a language were generalized to the language as a whole and considered as *real*. As most of the time they did not have the possibility to test invented examples on real-use data other than their internalized knowledge of the language, it was generally accepted that *possible* usage meant *real* usage.

Corpus linguistics has changed this tradition. Thanks to the analysis of billions of sentences of real-use language, we now know that languages are not deterministic systems, but rather they should be thought to be "probabilistic, analogical, preferential systems" (Hanks, 2013, p. 310). This groundbreaking discovery called for a review of many previous linguistic theoretical formulations and earlier established assumptions. New advances in computer science merged with huge quantities of digital material, including historical datasets, have allowed modern empirical linguists to study how people use words to communicate much more rapidly and efficiently than ever before. Today, the unprecedented amounts of naturally occurring language data have originated linguistic subfields such as corpus-based historical pragmatics and semantics and computational sociolinguistics. These disciplinary developments not only have yielded a deeper understanding of what meaning is, but also pointed at novel ways to study changes of how word meaning is constructed over time.

If certain twentieth century linguistics hypotheses have been challenged by real-use data, others have been later tested and confirmed. One such hypothesis is Firth's (1957, p. 11) famous intuition that "You shall know a word by the company it keeps." This intuition provided the foundation for his work on collocational meaning, which acknowledges the relevance of collocation in determining meaning. Collocational meaning has substantially contributed to the field of distributional semantics, the field of study concerned with measuring and categorizing how words are used based on patterns of usage. The core idea of distributional semantics is that meanings do not exist in isolation, but rather that words that are used and occur in the same contexts tend to purport similar meanings (Harris, 1954, p. 156). The distributional hypothesis is still central to most lines of inquiry of NLP techniques and has been applied to computational word vector models, including, for instance, the word2vec algorithm of Google. What is perhaps even more relevant to work on semantic change is the firm rejection of the belief that words can have a one-to-one relationship with meaning. As Harris argued (emphasis added), "We cannot say that each morpheme or word has a single or central meaning, or even that it has a *continuous* or coherent range of meanings" (Harris, 1954, p. 152). According to this claim, then, words in isolation do not possess meaning. We can only entail meaning from context; therefore, we can only detect changes in a word's meaning by analyzing patterns of changes in the word's context.

Another linguistic field that examines the relationship between language and meaning is cognitive linguistics. This field sees language as a mental phenomenon. Accordingly, it studies language as a window on the conceptual structure of the mind and considers how the evolution of language reveals changes in the common mindset over time. One of the core principles of cognitive linguistics is that meaning involves conceptualization, i.e., *construal*. Therefore, the way language is used informs us about the construction of meaning. As Croft (2009, p. 397) puts it: "The framing of an experience through the choice of a lexical item is a matter of construal." However, Croft himself has criticized traditional cognitive linguistics for considering language exclusively as a constellation of mental structures and processes. He argues that a comprehensive approach to language must take the fundamental function of language into account: communication. In other words, the interactive and the social dimensions of language, he claims, must be integrated in cognitive linguistics approaches to generate "a more general social–interactional model of language" (Croft, 2009). To do so, he argues, theories of pragmatics and sociolinguistics must be incorporated into cognitive linguistics.

This understanding of meaning in terms of its "discourse function" is particularly relevant to the line of research presented here. Integrating the dimension of discourse within cognitive linguistics means bringing a sociointeractional perspective to the construal of meaning, which serves "the purpose of communication" (Croft, 2009, p. 410). In agreement with Croft, our study starts from the conviction that the communicative property is essential to language and that discourse is indeed a crucial component of understanding the correlation between language, meaning, and society. Collective discourse represents the common, shared knowledge without which no communication would ever be possible. Focusing on either of the three aspects without considering their wider discourse embedding would be too restrictive and would yield only partial insights.

The approaches and theoretical frameworks discussed here, despite their differences in perspectives and goals, all agree that in the same way that meanings do not exist in isolation, discourses are not isolated entities, and that the interactive, pragmatic function of language must be considered. In our view, however, it is CDA that provides the best suited principles, theories, and methods to study language, meaning, and communication at the level of discourse. This is motivated by at least three arguments. First, critical discourse scholars define discourse as a form of social practice (Wodak and Fairclough, 1997, p. 258). Understanding discourse as a social phenomenon rather than a purely linguistic one (or mental one) entails that meaning is continually negotiated through interaction. Unlike semanticists who are concerned with the conventional meaning of words and sentences, critical discourse scholars are interested in understanding meaning as it is constructed during communication and for the purpose of communication. For critical discourse scholars, the goal is not to categorize conventional meaning, but rather to understand meaning as "socially constructed" through sign systems such as language. Discourse is in this way seen as "historically and culturally situated" rather than "eternal, absolute, and essential" (Locke, 2004, p. 11) and in a dialectical relationship with society: a discursive event shapes and is shaped by the situation, institution, and social structure that frames it (Wodak and Fairclough, 1997, p. 258).

Second, unlike conceptual historians, CDA discards a deterministic relation between discourses and society. As discourses are produced for specific purposes, their analysis entails a theorization and description not only of the social processes and structures that led to its production but also of the social mechanisms underpinning the way in which individuals or groups as social historical subjects create meanings through discourses. Therefore, because discourse is situated in time and space, CDA is particularly effective at uncovering the discursive nature of social and cultural changes (Wodak and Meyer, 2001, p. 7). This is especially true when analyzing public and media discourses, which consequently are the perfect avenues for CDA inquiry. In this respect, Wodak and Meyer (2001) notice how newsmakers often present themselves as neutral carriers of news who, supposedly through unbiased language, merely show issues of societal relevance to the public. On the contrary, CDA of media discourse has repeatedly highlighted the fundamental role of mass media in shaping meaning and discourse (*cfr. The Role of Media in the Construction of Migration Discourse*). Thus, as media are active coproducers of discourses, discourses produced via media similarly determine reality (Jäger, 2001, p. 36). According to this view (emphasis added), "discourse analysis is not (only) about interpretations of something that already exists [...] but about the analysis of the *production of reality* which is performed by discourse—conveyed by active people" (Jäger, 2001).

Third, CDA provides useful methods such as the DHA that applies a sociopragmatic, historical perspective to the theory of CDA allowing to assess the historical context in which topics are formulated and discussed. In this way, the construction of meaning through language use is studied in its full sociohistorical context and as a reflection of cultural values and political ideologies. In this sense, DHA provides a triangulation of linguistic, social, and historical data that cognitive linguistics cannot offer and that is currently missing in computational semantics, unless a social cognition level is integrated into the model or a cultural one (see, for instance, Hamilton et al., 2016). In *Methodology and Dataset*, we will describe DHA in more detail.

Computational Approaches to the Study of Language and Meaning

As it has already been said, the availability of large textual corpora and advances in computational semantics have prompted a wave of publications aimed at mapping changes in lexical semantics using distributional methods, particularly prediction-based word embedding models. This section reviews only a handful of the most recent and relevant studies, and it is by no means meant to be exhaustive[3].

Research on detecting semantic shifts of words typically divides large historical textual corpora in time periods or "word epochs" (Mihalcea and Nastase, 2012; Popescu and Strapparava, 2014) and identifies change in the context of the word, believed to have undergone a shift by measuring co-occurring words. The Google Books Ngrams corpus proved good correlation with human judgment for detecting differences in word usage and meaning over time (Gulordava and Baroni, 2011; Kim et al., 2014; Mitra et al., 2015). In terms of sociocultural semantic shifts, a number of studies have shown that smaller time spans are more suitable, whereas longer spans should be used to study more structural, linguistic shifts. For instance, Kim et al. (2014) and Liao and Cheng (2016) used a 1 year time span dataset, whereas Kulkarni et al. (2014) applied a granularity of 1 month. These works showed the value of computational methods to trace semantic shifts with time spans of less than a decade with a particular focus on cultural drift.

Distributional word representations attempt to capture more subtle changes that may not be identified by mere word frequencies. With this technique, meaning is represented with sparse or dense (embedding) vectors, produced from continuous lexical representations of word co-occurrence counts. A number of recent publications have shown that distributional word representations (Turney and Pantel, 2010; Baroni et al., 2014) provide an efficient way to perform these tasks. Although these models still use word frequencies as data source, the information is condensed into continuous lexical representations, a technique that proved to outperform the frequency-based methods in detecting semantic shifts (Kulkarni et al., 2014).

To compare word vectors across different models Kim et al. (2014) propose the *incrementally updated diachronic embedding models*, which allow to calculate cosine similarities directly between the same word in different time period models. The technique trains a model on the diachronically first-time slice and then it updates it with the data from the successive time periods, and it saves its state each time. The intuition behind this approach is that all these models are inherently related to each other and therefore comparable.

THE ROLE OF MEDIA IN THE CONSTRUCTION OF MIGRATION DISCOURSE

There is urgent consideration for understanding how the migration debate unfolds in the media. Indeed, research on the topic has been conducted in a large variety of fields (e.g., media studies, discourse studies, political communication studies, to name but a few), about myriad groups, across multiple public discourse scenarios, and over the most disparate time periods (Migration Observatory at the University of Oxford, 2016). These studies have consistently demonstrated that media play a crucial role in framing, indicating that public opinion about migration is largely informed by mass media[4]. For instance, by discussing migrants in a negative way as delinquents or criminals, media may trigger a "cultivation effect," which slowly changes readers' perception of reality (Arendt, 2010; Balabanova and Balch, 2010; Balch and Balabanova, 2016).

Media also play an agenda-setting role, for instance, in discussing migration in the context of welfare, economy, or security (Buchanan et al., 2003; Moses, 2006, p. 137–43; Eberl et al., 2018). Multiple studies, for instance, have demonstrated how American and European media have framed migrants in a negative way, either by emphasizing the dichotomy of "us" vs. "them" or by creating an urgency of crisis (Cottle, 2000; Cisneros, 2008; Arcimaviciene and Baglama, 2018; Eberl et al., 2018; Viola and Musolff, 2019). Similarly, with reference to the UK migration debate, it has been argued that "the media are active agents in developing immigration policy" (Threadgold, 2009). Research on the more recent phase of the migration debate in the United Kingdom indicates that public perception does not match the quantitative, economic, fiscal, and cultural realities of migration (Duffy and Frere-Smith, 2014); although the global migration rates have not changed dramatically over the past half a century, on the whole in many Western nations, "the political salience of migration has strongly increased" (Lucassen et al., 2010, p. 1–4). Migration advocacy organizations and nongovernmental organizations argue that the public discussion—rather than the actual facts—plays a crucial role in creating political positions and in informing policy priorities and government choices (Sharry, 2000; Spencer, 2011;

[3] As a full review would have been beyond the scope of this article, we refer the reader to two recent excellent surveys carried out by Kutuzov et al. (2018) and Tahmasebi et al. (2018).

[4] For a recent account of the relationship between migration and media, see, for instance (Viola and Musolff, 2019).

Katwala and Somerville, 2016), thus underlining the urgency of understanding how collective meaning is constructed around the migration debate.

Although UK media have generally discussed migration in negative and even dehumanizing terms (Musolff, 2015), research has also shown considerable fluctuation over time in the specific connotations, as narratives and counternarratives compete with each other in the public arena (Duffy and Frere-Smith, 2014; Burscher et al., 2015; Blinder and Allen, 2016). Race, for instance, has been discussed as a dominant context in the 1970s (Hartmann and Husband, 1974; Messer et al., 2012; Hall et al., 2013), whereas security issues started to emerge in the beginning of the twenty first century (Abbas, 2019). The variety of different voices in a wide range of media has created a complexity that is not easily resolved by close reading. Because manual content analysis is time-consuming, most of these studies have focused on a time span of a few years or a decade at most. Overall, compilations of a long-term perspective on the migration debate in UK media is still missing. For this reason, it has been argued that the field of migration studies "must move beyond thick description, single case studies, and quantification to address a set of more focused themes and questions." (Baker et al., 2008; Gabrielatos and Baker, 2008). In a recent study (Gabrielatos and Baker, 2008), corpus analysis, i.e., collocation and word frequency, has been applied to the discursive constructions of refugees and asylum seekers in a 140-million-word corpus of UK press articles spanning one decade (1996–2015). Their analysis indeed confirmed the "media confusion and conflation of definitions" of key terms.

These works underline the promise of identifying *patterns* in the discourse over longer periods of time, currently hidden in large amounts of digital data. Our study adds to this line of inquiry and aims to achieve such goal. By combining computational, linguistic, and historical approaches, the intention is to uncover the way meaning is constructed around the urgent theme of migration in public discourse. This may also add a more quantitative perspective to migration studies, in which big quantities of data are increasingly playing a fundamental role (Pisarevskaya et al., 2019).

METHODOLOGY AND DATASET

This study proposes an experimental method that merges two techniques to trace and understand the historical evolution of media discourse. It uses word vector models trained on the TDA with a 10-year sliding window (Kenter et al., 2015; Martinez-Ortiz et al., 2016) to identify changes in the contexts in which the words *migration*, *immigration*, and *emigration* were used over time. We then use DHA to triangulate the word2vec results with social and historical data in order to provide plausible clarifications of the processes underpinning such changes. With this experimental method, our goal is to widen the focus of the analysis from identifying that a word has changed its meaning over time to exploring plausible extralinguistic factors that may help us understand the elements at play in the construction of collective meaning.

The word2vec models were trained on a 10 year slice with a 2 year sliding window (e.g., one model from 1900 to 1910 and sliding windows of 1900–1902, 1901–1903, 1902–1904, and so on). The word embeddings were generated with CBOW models, with 100 dimensions, a window size of 5, and minimum word count of 5, and used five negative samples. Starting from a *seed term*, the obtained terms result from the semantic network of each semantic model constructed from the documents in each time window (Kenter et al., 2015). For example, if the selected time window is 1900–1910, the outputted terms will be aggregated from the semantic models of 1900–1902, 1901–1903, 1902–1904, and so on. The weight is calculated by a Gaussian distribution where the mean of the distribution is the center of the period with a standard deviation of 1:0. The intuition behind this is that the central years in a period are the most likely to semantically reflect that period due to echoes from the preceding years and anticipations of the next years.

We compute similarities for three seed terms, *migration*, *immigration*, and *emigration*, in different time periods. Because the models overlap, most of the semantic relations between words remain stable, allowing us to detect granular changes over the years. As it has already been argued in the literature, this method offers a way to understand gradually changing words that are used to articulate the same topic, concept, or idea (Hamilton et al., 2016), which, in turn, allows us to trace historical changes in the construction of meaning over longer periods of time.

We use DHA to analyze the obtained similarities. The DHA applies a historical dimension to the theory of CDA (van Dijk, 1997) as it intends context as essentially historical. The historical orientation permits the reconstruction of how texts and discourses are linked intertextually and interdiscursively over time. In other words, the method considers how texts and discourses are linked to each other, both in the past and in the present, as well as to extralinguistic social/sociological variables in an effort to diachronically reconstruct and explain discursive change (Reisigl and Wodak, 2009, p. 95, 120). In practice, this is done by triangulating linguistic, social, and historical data with the aim of understanding language use as a reflection of its cultural values and political ideologies. Resulting in a quasi-kaleidoscopic investigation, this principle of triangulation arguably minimizes the risk of biases, an aspect of CDA often criticized in literature as being a method highly dependent on the researcher's interpretation. Furthermore, we argue for two additional advantages of employing the proposed integrated method: first, being supported by enormous quantities of language data such as the TDA, the combination of DHA with word vector models further contributes to reduce the risk of biases, and second, it overcomes the limitation of looking at individual documents, typical of CDA studies. Although at this stage no CDA is performed on textual excerpts, DHA is still very useful in the task of interpreting the results and explaining larger patterns.

As for the dataset, the TDA archive contains every page of every issue of the newspaper from 1785 to 2013 for a total of

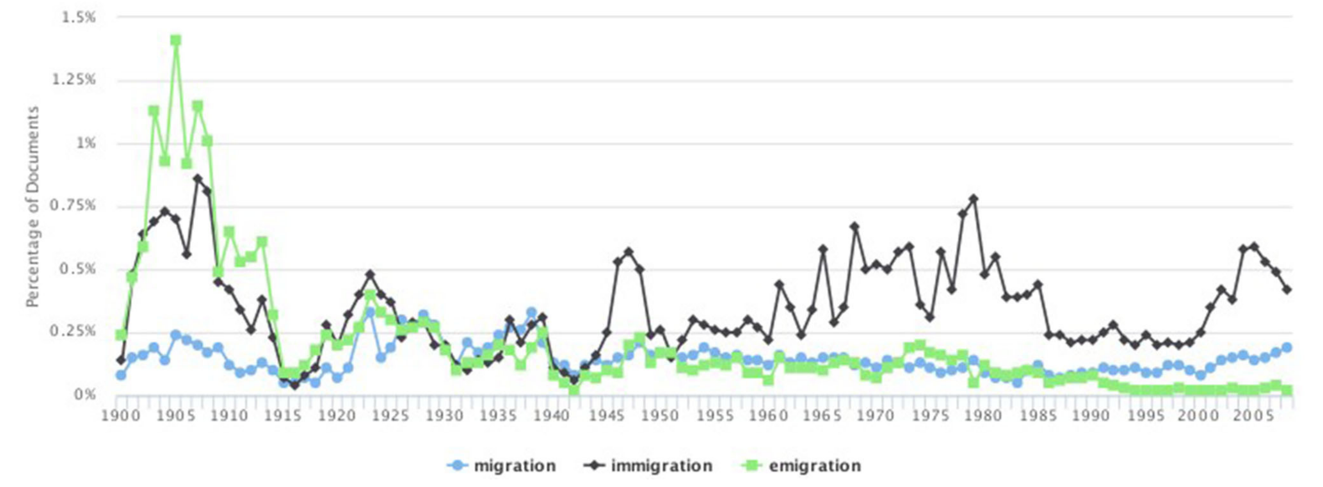


FIGURE 1 | Percentage of total documents by year containing the terms *migration*, *immigration*, and *emigration*.

more than 1.6 million pages from 70,000 issues, subdivided or zoned into 11.8 million articles, cataloged by category, including advertising, editorial and commentary, news, business, news, people, and photojournalism. The subset that we used (i.e., 1900–2000) contains 5,709,334,307 tokens (i.e., all the words) and 359,351,482 words (i.e., word types).

ANALYSIS

In this section, we demonstrate how the proposed method is applied by integrating DHA into the analysis of the word vector results. We have separated the discussion into five time periods following the quantitative results shown in **Figure 1**: 1900–1910, 1920–1930, 1945–1955, 1955–1985, and 1985–2000. The aim is not to explain historical changes or to confront historical knowledge with the computationally generated models, but to assess the value of including discourse-historical information into a computational model toward a greater understanding of the construction of meaning.

Preliminary Data Exploration

The data were first explored by graphing the frequency of occurrence of the seed terms *migration*, *immigration*, and *emigration* in the timeframe of reference. This initial step was performed to obtain insights of patterns and continuities within the British discourse around migration in the long century. **Figure 1** shows the percentage of the total documents by year that contains the terms of reference.

The graph yields interesting results. The first observation to be made concerns the fact that at the beginning of the century the three terms are sharply separated with a predominance of articles discussing *emigration*. This may indicate (1) that the migration discourse was not a generic topic incorporating the different aspects of human movement and (2) that emigration from Britain was discussed more often than immigration to the country and far more than migration in general. The second observation concerns the fact that between 1915 and 1940 this trend changes and the three topics seem to merge into one discourse. It is from 1945 that the current discourse polarization toward *immigration* can be first observed, whereas *migration* and *emigration* continue to converge until 1990, when they diverge again and *emigration* significantly decreases. This visualization of the frequency of the articles in *The Times* discussing the three topics already provides useful insights into how the wider migration discourse has developed in Britain: emigration, which used to dominate the discourse at the beginning of the century, has become, at the end of the observed period, the least frequently discussed topic. At the same time, in more recent times, the discourse has shifted almost completely toward immigration showing a clear change in the media construction of the migration discourse.

The graph is useful also to identify five main time slots or "word epochs," which are marked by the spikes of the three terms' frequencies. The first spike is that of *emigration* and *immigration* at the beginning of the century (1905–1910); the second spike can be noticed between 1920 and 1930 when the three terms are used almost equally. It is in 1945, just after World War II (WWII), that the graph shows a sharp rise in *immigration* (third spike); between 1955 and 1985, although with some fluctuations, *immigration* keeps prevailing over the other terms (fourth spike), and finally, between 1985 and 2000s, the graph shows a significant decrease of the term *emigration*.

This initial exploration of the dataset offers useful starting points, which will be investigated more in-depth in the next stage of the study. Thanks to the word vector analysis, we will be able to look more in detail into the semantic space of the three terms at the times when the spikes occur. At the same time, the DHA triangulation will clarify how such changes in the frequencies may be understood in relation to the concurrent sociohistorical events.

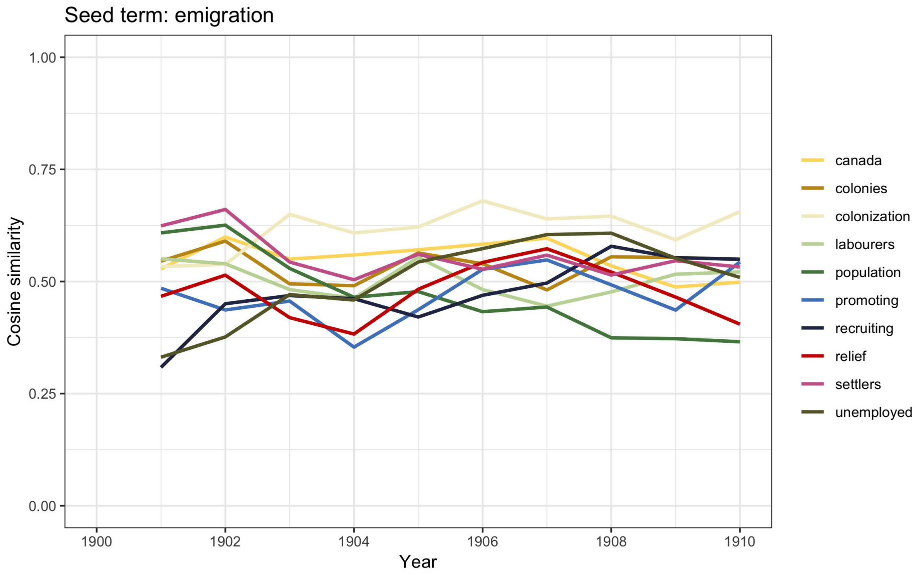


FIGURE 2 | Similarity scores for the 10 most frequent word vectors in TDA for *emigration*, 1900–1910.

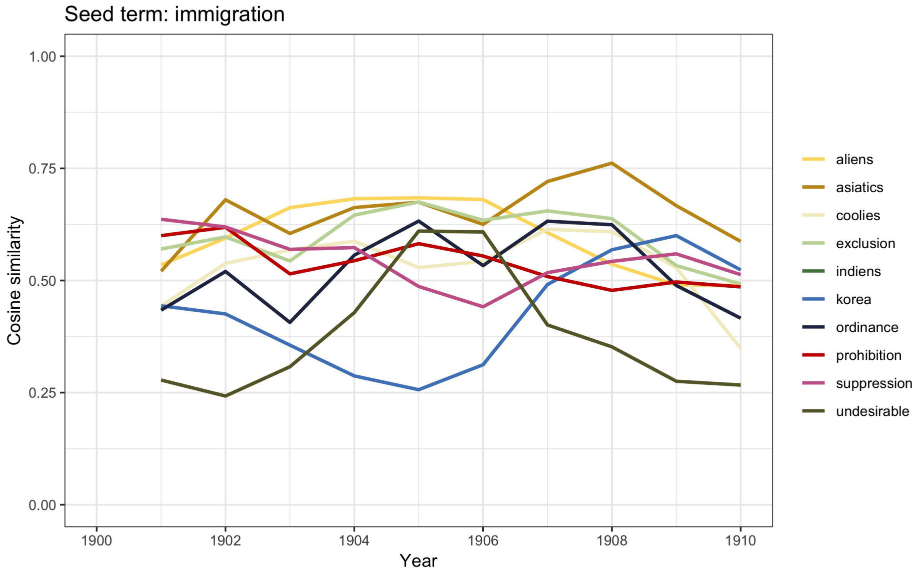


FIGURE 3 | Similarity scores for the 10 most frequent word vectors in TDA for *immigration*, 1900–1910.

1900–1910

During the 1900s and early 1910s, British emigration was at its highest: it is calculated that in England and Wales as many as 8.7 per thousand and, in Scotland, 18.7 per thousand emigrated (Bueltmann et al., 2012), placing Britain among the European countries with the highest emigration rates[5]. It is therefore not surprising that in those years emigration had become a topic discussed in political and public debates. The graph in **Figure 1** has already shown a peak in the number of documents containing the word *emigration*; **Figure 2** visualizes the similarity scores for the 10 most closely related terms to *emigration* from 1900 to 1910.

Population growth and industrialization were the main factors for emigration. Industries such as the mining and textile had been severely affected by industrialization, and many workers had to face sudden unemployment; emigration was often seen as an obvious solution. The word vector similarities show words such as *population* and *unemployed*, which may suggest that emigration was often framed in the press as the easy

[5] http://www.european-emigration.com/uk/

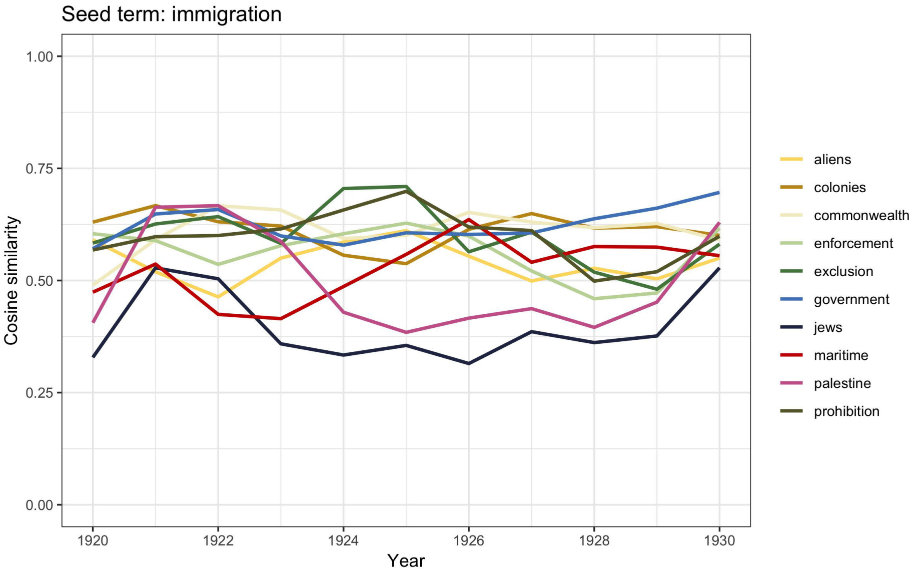


FIGURE 4 | Similarity scores for the 10 most frequent word vectors in TDA for *immigration*, 1920–1930.

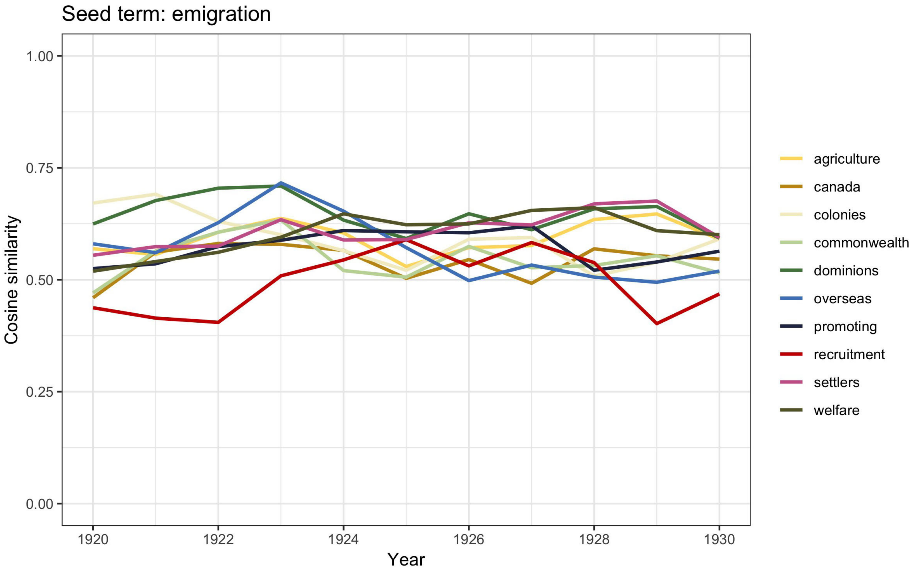


FIGURE 5 | Similarity scores for the 10 most frequent word vectors in TDA for *emigration*, 1920–1930.

answer to overpopulation and unemployment. It has already been noticed in the literature how the discussion would also sometimes incorporate imperial arguments according to which emigration was an efficient way to strengthen Britain's underpopulated colonies (Bueltmann et al., 2012). This would explain the presence of terms such as *colonization*, *colonies*, *settlers*, *recruiting*, and *promoting*.

The word vector similarities also show the word *Canada*. A DHA triangulation provides a potential explanation for this finding: between 1896 and 1914, Canada experienced rapid economic growth and development, thus becoming an attractive immigration destination. Estimates calculate that around 3 million immigrants arrived to Canada in those years, of which approximately one-third arrived from Britain (Lloyd, 2012, p. 137). Among other ethnic groups, British migrants were favored for several reasons: it was believed, for instance, that British immigrants would integrate more easily in Canada, as many Canadians already identified themselves as British. It was also believed that if British citizens had not moved to Canada in significant numbers, then Canada would be populated by "inferior" immigrants (Lloyd, 2012).

The graph in **Figure 1** also shows a relatively high number of documents containing the word *immigration* between 1900 and 1910. If Britain's industrial boom was a cause for emigration, in the century before, it had also attracted hundreds of thousands of immigrants. **Figure 3** visualizes the word vector similarity scores for *immigration*.

The analysis yields interesting results. The presence of words such as *asiatics*, *korea*, and *indiens* is not surprising: the opening of the Suez Canal in 1869 had facilitated immigration from India and China. However, the finding acquires a specific connotation if we look at the other terms: *exclusion*, *suppression*, *undesirable*, and *coolies*. By the end of the nineteenth century, anti-immigrant feelings were on the rise, and calls for immigration control laws became more and more pressing; the presence of terms with a stronger *relationship measure* (RM), i.e., the weight such as *prohibition*, *ordinance*, and *suppression* suggests such tensions. The xenophobic sentiments are apparent not only in the terms *undesirable* and *exclusion* but also in *aliens*, which most likely refers to the Aliens Act, entered into force in 1905. Finally, another important observation to be made concerns the absence of reference to the three largest groups of immigrants who had arrived to Britain in those years from Germany, Ireland, and southern Italy. This absence may suggest that at this point in history the hostility was mainly directed toward Asian immigrants.

1920–1930

The graph in **Figure 1** shows how in the period between the end of WWI and 1930 the discourse incorporated all the three terms almost equally; we therefore computed word2vec similarities for all the three keywords. **Figure 4** shows the results for *immigration*, **Figure 5** displays the scores for *emigration*, whereas similarities for *migration* are visualized in **Figure 6**.

In those years, Britain was going through many social changes. Tension had arisen between white merchant seamen returning from war and migrant seamen, who in the meantime had replaced them. Violent riots and confrontations between the two opponents led to the proclamation of the 1925 Act which, by factually banning migrant seamen, became "the first instance of state-sanctioned race discrimination inside Britain to come to widespread notice" (Tabili, 1994, p. 56). In the time slot 1920–1930, the word vector similarities for *immigration* contain the words *maritime*, *exclusion*, *enforcement*, and *prohibition*. Meanwhile anti-Semitism was on the rise not only in Britain but also across Europe. Between 1882 and 1919, Jewish numbers in Britain had significantly increased from 46,000 to 250,000[6]; they were mostly escaping from Russia, where they were harshly discriminated. This may be related to the words *jews* and *Palestine* appearing in the similarities scores.

Emigration from Britain during the nineteenth and early twentieth century was primarily overseas; migration overseas was a major feature of Victorian society (Pooley and Turnbull, 1998, p. 258). It has been calculated that between 1840s and 1930s, people who emigrated overseas from Britain outnumbered those who migrated to Britain. The preferred emigration destinations were by far North America and Canada, but toward the beginning of the twentieth century, New Zealand, Australia, and South Africa had become more and more popular. This explains the presence of closely related words such as *Canada*, *colonies*, *commonwealth*, *overseas*, and *dominions*. The other words with high similarity scores are *promoting*, *recruiting*, and *agriculture*. Because of the greater distance and lesser knowledge of Australia and New Zealand, emigration to these countries was typically arranged by companies providing assisted passages. These companies would often recruit people whose sets of skills could benefit specific economic needs in the countries of destinations. This ultimately meant that these companies had control over the characteristics of those who moved (Richards, 1993; Haines, 1994; Pooley and Turnbull, 1998).

Although, traditionally, the majority of studies on race, ethnicity, and racism in Britain trace the beginning of xenophobic sentiments since 1945, some authors (e.g., Solomos, 2003) have stated that, in fact, it was during the interwar period that the question of racial difference started to enter the political debate of immigration. The social decay, particularly of seaport towns, started to be associated with black communities; shipping industry trade unions capitalized on this discriminatory belief and campaigned in order to restrict employment to white seamen. As we have already said, these discriminatory actions led to the 1925 Act; however, additional practices were reinforced, such as legalized different rates of pay based on race (Hepple, 1983, p. 44–45; Joshua et al., 1983), which were meant to prevent British citizens of a different race from settling in the country. But there was also another concern related to their settlement, and it had to do with the fear of a "mixed race" population as a result of mixed race unions (Rich, 1986, p. 120–44; Ramdin, 1999, 2017). The word vector similarity scores for *migration* reported in **Figure 5** report the words *racial*, *conflicts*, *epidemics*, *immigrants*, *culture*, and *segregation*, which can be understood in the light of this historical contextualization. Because of the conflicts that would occur in some of the port towns, as well as the spread of an image of black people as sources of social problems, these communities were labeled as "aliens" and perceived as threatening to British culture (Solomos, 2003, p. 47). This set up the foundational arguments that characterized the political debate on "colored immigrants" following WWII.

1945–1955

We will now move to analyze the third period between 1945 and 1955. The graph in **Figure 1** shows that after WWII there was a peak in the number of documents discussing *immigration*, whereas *migration* and *emigration* showed a similar decreasing trend. As the similarity scores for *emigration* are very similar to those analyzed in the previous period, to avoid repetition in this section we comment the results for immigration (**Figure 7**). The full computed similarity scores are provided in the **Supplementary Material**.

The migration debate changed quite dramatically after WWII as more than 11 million people became displaced from their home countries throughout Europe, including former prisoners of war, released forced laborers, and survivors of concentration camps. The United Kingdom experienced

[6] https://www.bod.org.uk/jewish-facts-info/jews-in-britain-timeline/

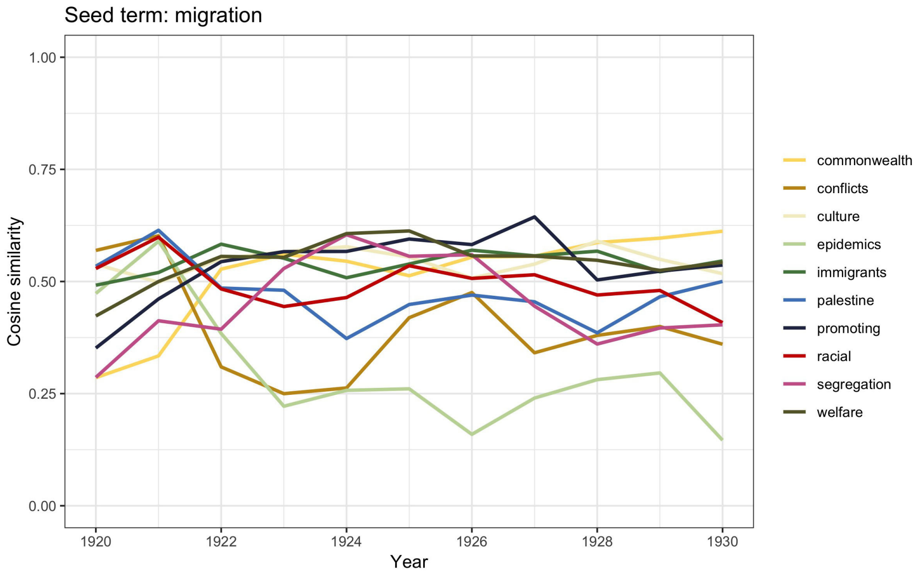


FIGURE 6 | Similarity scores for the 10 most frequent word vectors in TDA for *migration*, 1920–1930.

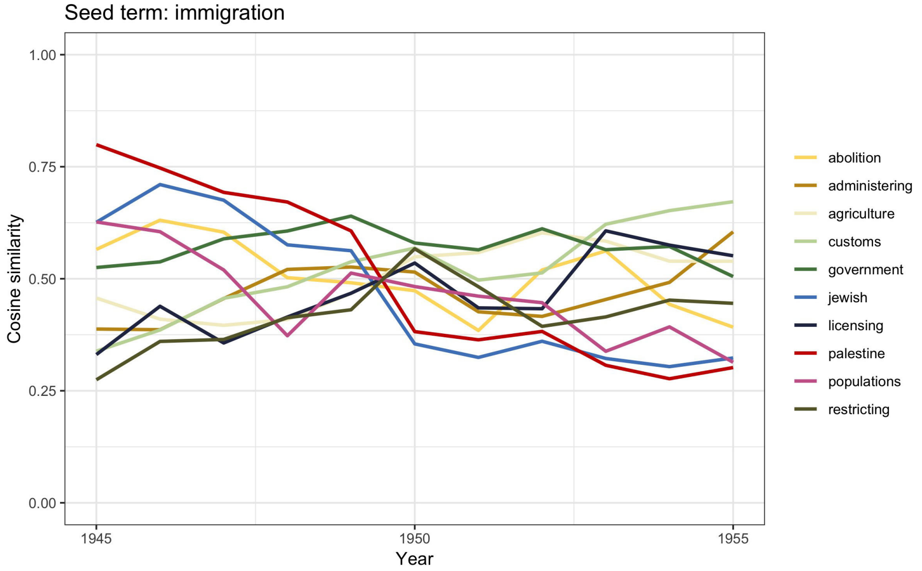


FIGURE 7 | Similarity scores for the 10 most frequent word vectors in TDA for *immigration*, 1945–1955.

considerable immigration from such displaced persons and other refugees from Europe. After 1945, the topic of immigration became connected with the persecution of the European Jews as the urgency of creating a new homeland for Jews in Palestine, debated throughout the war, had become a matter of public debate. This is visible in the word vectors similarity scores for *immigration*: *palestine* and *jewish*, which reflects the illegal immigration of Jews into Mandatory Palestine, governed by the United Kingdom between 1920 and 1948 (Kochavi, 1998; El-Eini, 2006; Cohen, 2014). It is interesting to notice that the term *palestine* significantly decreases after 1950 when the creation of Israel ended the British governmental involvement with Palestine. It is also worthwhile mentioning that the largest group to enter the United Kingdom between 1945 and 1954—the almost 1,000,000 Irish migrants—is not visible in the results. This would be in line with claims made by historians such as Solomos (2003, p. 42) who have highlighted how Irish immigrants hardly left a trace in public debate. Neither do we see references to Polish immigration, even though the secondary literature tells us that about 150,000 Polish army veterans were resettled in

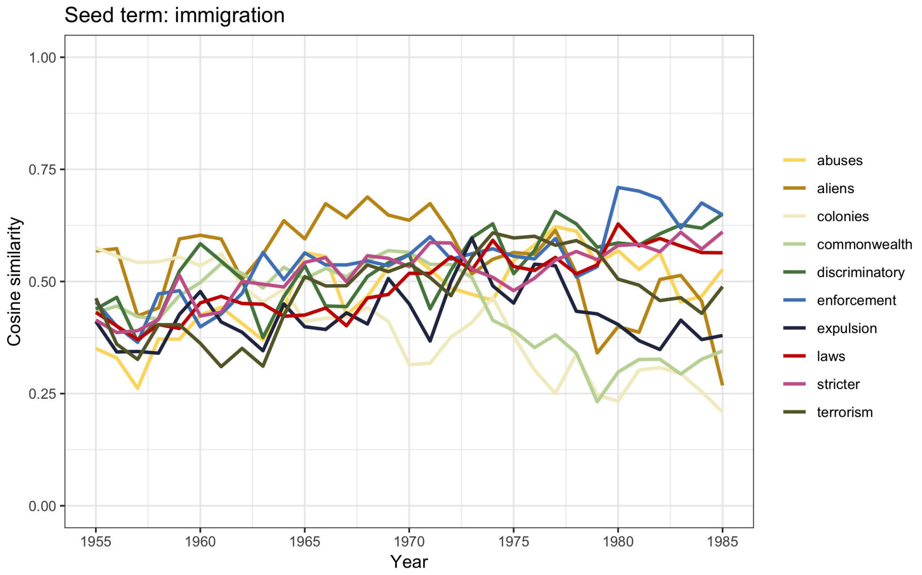


FIGURE 8 | Similarity scores for the 10 most frequent word vectors in TDA for *immigration*, 1955–1985.

the United Kingdom between 1946 and 1949 (Sword, 1986; Blaszczyk, 2017).

The other terms appearing in **Figure 7**—*government*, *licensing*, *administering*, *restricting*, *abolitions*—may be understood in the context of the racialization of the political debate of immigration after the 1945s. Those are the years when the British government tried in many ways to prevent black migrants from entering Britain (Joshua et al., 1983; Carter et al., 1987), even when they were British. For instance, in 1948, the British Nationality Act distinguished between British subjects who were citizens of the United Kingdom and its colonies and those who were Commonwealth citizens, even though the right to enter and live in Britain was granted to both categories (Evans, 1983, p. 59–61; Evans, 1986). A number of nontransparent measures were additionally adopted to impede black immigration as much as possible (Carter et al., 1987). These actions mirror the sharp contrast between the government's intention of restricting the settlement of black colonial British citizens on the one hand and the wish to not undermine the notion of Britishness on the other (Joshua et al., 1983). According to Solomos (2003, p. 54) it was during this time that the political and media debate of immigration revolved heavily around race and color and on the effects that black immigration would have on the "racial character of the British people," the *customs*, the national identity, and "Britishness."

1955–1985

Throughout the 1950s, the British political and public debate of migration sharply polarized toward immigration, which was felt as a much more pressing issue than emigration. The graph in **Figure 1** shows a clear peak for articles discussing *immigration* between 1955 and 1985; *emigration* and *migration* on the contrary displayed a similar low frequency. We here again discuss the similarity scores for *immigration* as shown in **Figure 8**.

The immigration debate of those years focused on the need to control "colored" immigration and revolved around two main arguments both aiming to legitimize the need to limit the number of nonwhite immigrants in the country. One argument concerned the urgent need for regulating immigration and called for active governmental interventions; this argument may be visible in the words *stricter*, *laws*, *enforcement*, and *aliens.* The other—visible in the words *colonies*, *commonwealth*, and *terrorism*—concerned the alleged social urgency of crime, employment, and housing in relation to too many colored immigrants in the country (Solomos, 2003, p. 53). At the same time, however, arguments against the introduction of more rigid controls were also raised both by conservative and labor politicians. Although it is not entirely clear which motivations were brought into the discussions, it seems that at least to an extent these measures were accused to be mere discriminatory practices (i.e., *discriminatory*, *abuses*) and cause for embarrassment to Britain as head of the Commonwealth and colonies (i.e., *Commonwealth*, *colonies*).

Eventually, all these arguments led to the 1962 Commonwealth Immigrants Act, which may be seen as the government's attempt to implement a measure that appeared to control immigration in general, whereas the real intention was in fact to limit black immigration only. The direct consequence of this Act was a high politicization of the term "immigration" itself, which *de facto* became code for racism (Solomos, 2003, p. 56). The 1962 Commonwealth Immigrants Act had in this way set the terms for the beginning of a political process enforcing even stricter immigration measures, such as the 1971 Immigration Act. With this Act, the notion of *citizenship* distinguished between partial and non-partial citizens, the former being the

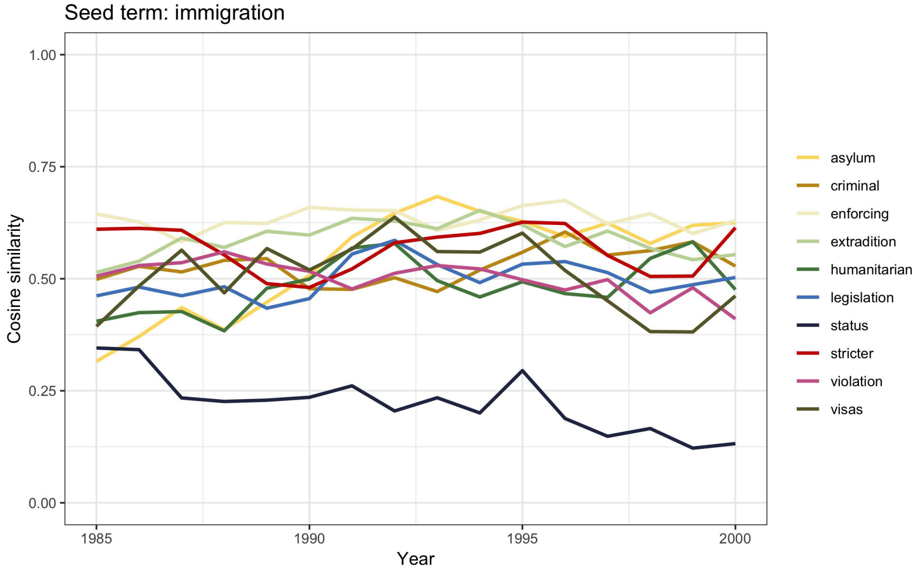


FIGURE 9 | Similarity scores for the 10 most frequent word vectors in TDA for *immigration*, 1985–2000.

only ones having the right to reside in Britain. Factually, this measure was institutionalizing racism as it allowed only white Commonwealth citizens to enter and settle in Britain. From 1979, the Thatcher administration further strengthened the controls on immigration from the Commonwealth, starting from passing the 1981 British Nationality Act. This Act was dividing British citizenship into three categories: British citizens, British Dependent Territories Citizens, and British Overseas Citizens. The last category affected most British citizens of Asian origin who became in this way deprived of their right to live in Britain. This and other measures were justified by the argument that stricter immigration regulations were necessary to limit the number of people having access to social resources and services. However, by doing so, issues such as employment, housing, education, and law and order became highly racialized (Hall et al., 2013). Consequently, the focus of attention in the British public and political debate about immigration shifted from the question of immigration *per se* toward the identification of race as the source of the problem (Castles et al., 1984; Macdonald and Toal, 2014).

1985–2000

This section focuses on the last part of the century; the identification of the general debate of migration with immigration that had started in 1945 continued to the end of the century. The graph in **Figure 1** reports an overall lower frequency of the number of articles discussing *migration*, *immigration*, and *emigration*, but it is once again *immigration* that shows the highest frequency. **Figure 9** shows the similarity scores for *immigration*.

Partly due to the dismemberment of the communist bloc in Eastern Europe, the public debate during the early 1990s was dominated by discussions on asylum seekers and refugees. Such discussions soon became closely intertwined with the political and public debate on immigration—visible in the words *asylum*, *visas*, and *status*. Similarly to the arguments used only a few years before to legitimize stricter immigration measures against non-white immigrants, the focus of the debate was once again on the “alarming” growing number of asylum seekers and refugees in the country (Spencer, 1994, 1997; Macdonald and Toal, 2014). Such growing number was seen as a major concern for which stricter regulations were urgently required (i.e., *stricter*, *legislation*, *enforcing*). The discussions eventually led to the enforcement of the 1993 Asylum and Immigration Appeals Act, which aimed to reduce the number of asylum seekers and refugees able to claim sanctuary.

The presence of the words *violation*, *extradition*, and *humanitarian* could be explained by the events that led to the 1999 Immigration and Asylum Act. In 1996, the European Court of Human Rights had intervened on a specific deportation trial ruling that it was a case of human rights violation. The court’s decision was effectively limiting the governments’ power of deporting subjects for matters of national security. The British government’s reaction was to set up a special commission that would work around the court’s decision, essentially preserving the government’s right to deport. The possibility to increase deportations together with harsher measures toward ethnic minorities and asylum seekers eventually constituted the core of the 1999 Immigration and Asylum Act.

In the last part of the century, the semantic connection between asylum seekers and refugees and the old concerns regarding the social and cultural dangers of immigration became tighter and tighter, reflecting the institutional legitimization of racial discrimination.

DISCUSSION

Unlike traditional distant reading approaches, the word vector similarity scores computed for 1900–2000 revealed trends and shifts that allowed for a wider contextualization and a much broader view than it would have been possible with smaller samples of data or analyses carried out over shorter periods of time. This demonstrates how computational distant reading reveals the *longue durée* of big history (Armitage and Guldi, 2014; van Eijnatten et al., 2014). Specifically, two macrotrends could be identified regarding the way immigration and emigration were discussed in the British public discourse. On the one hand, words such as *promoting*, *recruiting*, and *relief* were found associated with emigration *from* Britain suggesting that it was framed as a positive phenomenon, described as good not only for Britain but also for the emigration destinations of British citizens. The discourse-historical triangulation confirmed that emigration as the easy solution to overpopulation and unemployment, as well as an effective way to strengthen Britain's underpopulated colonies, was the main argument used to support this positive narrative. On the other hand, the opposite, yet consistent trend was found to be true for immigration *to* Britain, particularly from the colonies, which was consistently found in the context of negative terms such as *exclusion*, *undesired*, and *restricting*. Similarly to the way emigration was positively constructed through a range of legitimizing arguments, the construction of immigration as "negative for the country" was constructed through a variety of recurrent, yet powerful images: at times, it was associated with other social problems in the country (e.g., housing, unemployment), and at other times with the decay of British culture (e.g., loose customs) and with law and order issues (e.g., rise in crime, terrorism) or other general threats (e.g., invasion of immigrants, alarming numbers).

Another clear pattern was found within the semantic space of immigration and in the way in which, over the course of the century, immigration concerns became more and more associated with social categories of ethnic minorities (e.g., non-white seamen, Jews, immigrants from former colonies, asylum seekers), whereas larger groups of immigrants (e.g., Italians, Germans, Polish, Irish) were practically absent in the similarity scores. This could be an indicator of a process of racialization of the public debate around immigration, which, through a number of repeated arguments and narratives, targeted different categories of individuals at different times. The fact that immigration was embedded within widely debated issues of social urgency that required stricter laws and immediate intervention is also visible in the many terms referring to rules and regulations that were consistently present in the semantic space of *immigration* but that were totally missing in the similarity scores of *emigration*. This was found to be the case also when emigration from Britain had reached impressive figures, in fact, even when, historically, Britain was experiencing higher emigration than immigration. As emigration was seen as beneficial to the country, calls for more draconian measures referred exclusively to the immigration debate.

Word vector models are an extreme form of distant reading as the text structure itself disappears entirely; however, by integrating the technique with the discourse-historical triangulation, emerging larger patterns could be identified and understood. Although not optimal, the combination of very large quantities of data, the researcher's exclusion from the results, and the historical triangulation allows for a more empirical, reproducible, and comprehensive analysis that overcomes the limitations of either fully interpretative methods or conclusions drawn on fragmented data.

CONCLUSIONS

This article offered a methodological contribution to the field of computational sociolinguistics. We combined neural word embeddings and methods of CDA to study the historical construction of public meaning at the level of discourse. The study's foundational hypothesis was that, because meanings are not established in isolation but are socially constructed, the analysis should move from word level to discourse level. To add historical depth and a broader contextualization, we also argued for a computational diachronic approach. In order to do so, we used word vector models built according to sliding time windows of one decade each and analyzed public discourse about migration in twentieth century Britain.

Our contribution was innovative in at least three aspects. First, by choosing to focus on migration as a *topic* rather than as a word—as it is typical in word embedding studies—we operated within a linguistic framework of conceptual history. This meant that while word vector models allowed us to trace the different vocabularies used in specific discourses over time, the discourse-historical angle provided us with the necessary framework to understand the correlation between language, meaning, and society. Second, the diachronic analysis allowed us to make sense of the variations and continuities in the discourse. Specifically, the examination of the observed changes within the historical context showed a recursive cycle: historical events were reflected in the public discourse; this, in turn, shaped explicit meanings in the public debate, which contributed to trigger further historical events. Third, the study focused on a specific type of discourse, *media discourse*. Mass media both reflect and influence public discourse as they are the main vehicle of knowledge-circulation and opinion-formation. Because they influence the way topics and events are perceived, mass media also impact both the public and policy makers. Thus, understanding media coverage of specific topics is essential not only to understand the corresponding society's response, but also to comprehend political and public attitudes that shape behavior, policy, and, finally, language. In reflecting collective discourse, media represent the common, shared knowledge that makes communication possible.

The study of public discourse about migration in the United Kingdom from 1900 to 2000 deepened our understanding of how meaning is constructed in language over time and how it shapes and it is shaped by sociohistorical change. The analysis revealed significant shifts in both the frequency and essence of how the meanings attached to *migration*, *immigration*, and *emigration* were formulated and discursively constructed as resulting from changing historical concerns in British society. For instance, at the beginning of the century, emigration was frequently discussed in the public media and promoted as an obvious solution to overpopulation and unemployment. A few years later, it was framed as an imperial necessity, the only way to strengthen Britain's underpopulated and "wrongly" populated colonies and dominions. After the devastation of WWII, migration was associated with displaced persons and the search for a Jewish homeland, and from the 1950s on, race and immigration became the dominant context of the great "migration debate." Toward the end of the century, the dominant context of immigration as a publicly debated topic shifted again in the direction of internal social security and fear. These subtle shifts in the meanings attached to the migration discourse could be seen in the changing semantic space of the vectors and understood through the discourse-historical triangulation.

This research was based on the TDA. Although one of the most significant newspapers in the United Kingdom, *The (London) Times* is believed to have reflected English establishment, government, metropolitan interests, and empire. A fuller representation of public discourse would need to include other voices and, in the second half of the century, different media. Such a multimedia approach would offer a promising way to add a more comprehensive perspective to this line of inquiry. Nevertheless, despite its ideological, commercial, and political agendas, the collection remains an invaluable source for the study of public discourse. After all, in order to survive, newspapers must ultimately reflect contemporary debates of societal relevance.

Finally, the neural word embeddings used here should not be seen as a substitute for close reading strategies, rather as a complementary methodological approach that may allow researchers to adopt a zoom-out perspective to deal with large textual collections spanning over a long period of time. This macroperspective in combination with the discourse-historical triangulation may be used as a way to merge close and distant reading and proved effective at providing a comprehensive vision of the way meaning is socially constructed. By integrating these techniques, we also aimed to avoid confining the computational analysis to a role of support to critical analysis and to contribute to bridge the binary division between distant and close reading. This mixed method allows us to examine how language that is used to articulate public discourse is shaped by social changes and in turn may have helped to accelerate those changes. The study rested on the foundation that discourse conveys historical meanings. Therefore, understanding discourse changes means understanding social changes, and conversely, social changes will be reflected in changes in discourse. Although correlation does not prove direct causation, it is hoped that the method will highlight the importance of including sociohistorical data into a computational analysis so as to assist researchers to refine the quantitative results, make sense of them, and open avenues for understanding linguistic change.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

This article is the product of a collaboration between historical linguist LV and cultural historian JV. All authors performed the historical and computational analysis and contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors sincerely thank Jonathan de Bruin for his diligent help with data visualization.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.00064/full#supplementary-material

REFERENCES

Abbas, M. S. (2019). Conflating the muslim refugee and the terror suspect: responses to the syrian refugee "crisis" in brexit Britain. *Ethnic Racial Stud.* 42, 2450–2469. doi: 10.1080/01419870.2019.1588339

Arcimaviciene, L., and Baglama, S. H. (2018). Migration, metaphor and myth in media representations: the ideological dichotomy of "them" and "us". *SAGE Open* 8:215824401876865. doi: 10.1177/2158244018768657

Arendt, F. (2010). Cultivation effects of a newspaper on reality estimates and explicit and implicit attitudes. *J. Media Psychol.* 22, 147–159. doi: 10.1027/1864-1105/a000020

Armitage, D., and Guldi, J. (2014). The return of the *Longue Durée*: an Anglo-American perspective. *Ann. Hist. Sci. Soc.* 70, 219–247. doi: 10.1017/S2398568200001126

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press. *Discourse Soc.* 19, 273–306. doi: 10.1177/0957926508088962

Balabanova, E., and Balch, A. (2010). Sending and receiving: the ethical framing of Intra-EU migration in the European press. *Eur. J. Commun.* 25, 382–397. doi: 10.1177/0267323110381005

Balch, A., and Balabanova, E. (2016). Ethics, politics and migration: public debates on the free movement of romanians and bulgarians in the UK, 2006–2013. *Politics* 36, 19–35. doi: 10.1111/1467-9256.12082

Baroni, M., Dinu, G., and Kruszewski, G. (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Baltimore, MD: Association for Computational Linguistics), 238–247. doi: 10.3115/v1/P14-1023

Berry, D. M., and Fagerjord, A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge; Malden, MA: Polity.

Blaszczyk, A. (2017). The resettlement of polish refugees after the second world war. *Forc. Migr. Rev.* 54, 71–73. Available online at: https://search-proquest-com.proxy.library.uu.nl/docview/1876052308?accountid=14772

Blinder, S., and Allen, W. L. (2016). Constructing immigrants: portrayals of migrant groups in British national newspapers, 2010–2012. *Int. Migr. Rev.* 50, 3–40. doi: 10.1111/imre.12206

Brunner, O., Conze, W., and Koselleck, R. (Eds.). (1972). *Geschichtliche Grundbegriffe: Historisches Lexikon Zur Politisch-Sozialen Sprache in Deutschland*. Vol. 8. Stuttgart: E. Klett.

Buchanan, S., Grillo, B., and Threadgold, T. (2003). *What's the Story?: Results From Research Into Media Coverage of Refugees and Asylum Seekers in the UK*. London: Article 19.

Bueltmann, T., Gleeson, D., and MacRaild, D. M. (2012). *Locating the English Diaspora, 1500-2010*. Liverpool: Liverpool University Press. Available online at: http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=867097.

Burscher, B., van Spanje, J., and de Vreese, C. H. (2015). Owning the issues of crime and immigration: the relation between immigration and crime news and anti-immigrant voting in 11 countries. *Elect. Stud.* 38, 59–69. doi: 10.1016/j.electstud.2015.03.001

Carter, B., Harris, C., and Joshi, S. (1987). The 1951–55 conservative government and the racialization of black immigration. *Immigr. Minor.* 6, 335–347. doi: 10.1080/02619288.1987.9974665

Castles, S., Booth, H., and Wallace, T. (1984). *Here for Good: Western Europe's New Ethnic Minorities*. London: Pluto Press. Available online at: https://www.abebooks.co.uk/9780861047529/GOOD-WESTERN-EUROPES-NEW-ETHNIC-0861047524/plp

Cisneros, J. D. (2008). Contaminated communities: the metaphor of "immigrant as pollutant" in media representations of immigration. *Rhet. Public Affairs* 11, 569–601. doi: 10.1353/rap.0.0068

Cohen, M. J. (2014). *Britain's Moment in Palestine: Retrospect and Perspectives,* 1917–1948. *Israeli History, Politics and Society 55*. London: Routledge.

Cottle, S. (ed.). (2000). *Ethnic Minorities and the Media: Changing Cultural Boundaries*. Buckingham; Philadelphia, PA: Open University Press.

Couldry, N. (2008). Mediatization or mediation? Alternative understandings of the emergent space of digital storytelling. *New Med Soc.* 10, 373–391. doi: 10.1177/1461444808089414

Croft, W. A. (2009). *Toward a Social Cognitive Linguistics*. Amsterdam; Philadelphia, PA: John Benjamins Publishing Company. Available online at: https://benjamins.com/catalog/hcp.24.25cro

De Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights*. New York, NY: Fordham University Press.

De Bolla, P., Jones, E., Nulty, P., Recchia, G., and Regan, J. (2019). The conceptual foundations of the modern idea of government in the british eighteenth century: a distributional concept analysis. *Int. J. History Culture Modern*. 7:575. doi: 10.18352/hcm.575

Duffy, B., and Frere-Smith, T. (2014). *Perceptions and Reality: Public Attitudes to Immigration*. London: Ipsos More Social Research Institute. Available online at: https://www.ipsos.com/ipsos-mori/en-uk/perceptions-and-reality-public-attitudes-immigration

Eberl, J.-M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., et al. (2018). The European media discourse on immigration and its effects: a literature review. *Ann. Int. Commun. Assoc.* 42, 207–223. doi: 10.1080/23808985.2018.1497452

El-Eini, R. (2006). *Mandated Landscape: British Imperial Rule in Palestine,* 1929–1948. London: Routledge.

Evans, J. (1983). *Immigration Law, 2nd Edn*. Available online at: https://digitalcommons.osgoode.yorku.ca/faculty_books/250

Evans, J. M. (1986). The Development of British Immigration Law by V. Bevan. London: Croom Helm, 1986, xxxvii 388 (bibliography and index) 55 pp (hardback £29.95). *Legal Stud.* 6, 340–342. doi: 10.1017/S0261387500013805

Firth, J. R. (ed.). (1957). *Studies in Linguistic Analysis*. Oxford: Blackwell.

Gabrielatos, C., and Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *J. English Linguist.* 36, 5–38. doi: 10.1177/0075424207311247

Gavin, M. (2015). *The Arithmetic of Concepts: A Response to Peter de Bolla*. Modelling Literary History (blog). Available online at: http://modelingliteraryhistory.org/2015/09/18/the-arithmetic-of-concepts-a-response-to-peter-de-bolla/

Gooding, P., Terras, M., and Warwick, C. (2012). *The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book*. Vol. 28. Available online at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-myth-of-the-new-mass-digitization-distant-reading-and-the-future-of-the-book.1.html

Gooding, P., Terras, M., and Warwick, C. (2013). The myth of the new: mass digitization, distant reading, and the future of the book. *Literary Linguist. Comput.* 28, 629–639. doi: 10.1093/llc/fqt051

Graham, S., Milligan, I., and Weingart, S. B. (2016). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Gulordava, K., and Baroni, M. (2011). "A distributional similarity approach to the detection of semantic change in the google Books Ngram Corpus." in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (Edinburgh: Association for Computational Linguistics), 67–71. Available online at: https://www.aclweb.org/anthology/W11-2508

Haines, R. (1994). Indigent misfits or shrewd operators? government-assisted emigrants from the United Kingdom to Australia, 1831–1860. *Popul. Stud.* 48, 223–47. doi: 10.1080/0032472031000147776

Hall, S., Critcher, C., Jefferson, T., Clarke, J., and Roberts, B. (2013). *Policing the Crisis: Mugging, the State, and Law and Order*, 2nd Edn. Basingstoke; Hampshire: Palgrave Macmillan.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin: Association for Computational Linguistics), 1489–1501. doi: 10.18653/v1/P16-1141

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press. Available online at: http://site.ebrary.com/id/10651991

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162.

Hartmann, P. G., and Husband, C. (1974). *Racism and the Mass Media: A Study of the Role of the Mass Media in the Formation of White Beliefs and Attitudes in Britain*. London: Davis-Poynter.

Hepple, B. A. (1983). Judging equal rights. *Curr. Leg. Probl.* 36, 71–90. doi: 10.1093/clp/36.1.71

Jäger, S. (2001). "Discourse and knowledge: theoretical and methodological aspects of a critical discourse and dispositive analysis," in *Methods of Critical Discourse Analysis*, eds R. Wodak and M. Meyer (London: Sage), 32–62.

Jockers, M. L. (2013). "Macroanalysis: digital methods and literary history," in *Topics in the Digital Humanities.* Urbana, IL: University of Illinois Press.

Joshua, H., Wallace, T., and Booth, H. (1983). *To Ride the Storm : The 1980. Bristol Riot and the State*. London: Heinemann.

Katwala, S., and Somerville, W. (2016). *Engaging the Anxious Middle on Immigration Reform: Evidence from the UK Debate*. Washington, DC: Migration Policy Institute.

Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). "Ad hoc monitoring of vocabulary shifts over time," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM'15* (Melbourne, VIC: ACM Press), 1191–1200. doi: 10.1145/2806416.2806474

Kim, Y., Chiu, Y. -I., Hanaki, K., Hegde, D., and Petrov, S. (2014). "Temporal analysis of language through neural language models," in *Proceedings of the ACL 2014. Workshop on Language Technologies and Computational Social Science*, 61–65.

Kochavi, A. J. (1998). The struggle against Jewish immigration to Palestine. *Middle East. Stud.* 34, 146–167. doi: 10.1080/00263209808701236

Koselleck, R. (1992). "Lexikalischer rückblick," in *Geschichtliche Grundbegriffe: Historisches Lexikon Zur Politisch-Sozialen Sprache in Deutschland*, Vol. 7, eds O. Brunner, W. Conze, and R. Koselleck (Stuttgart: Klett-Cotta), 380–389.

Koselleck, R. (2004). *Futures Past: On the Semantics of Historical Time*. New York, NY: Columbia University Press.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2014). Statistically significant detection of linguistic change. *arXiv:1411.3315 [Cs]*.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. *arXiv [Preprint] arXiv:1806.03537 [Cs]*, June. Available online at: http://arxiv.org/abs/1806.03537

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lee, C. (2019). How are 'immigrant workers' represented in Korean news reporting?—A text mining approach to critical discourse analysis. *Dig. Schol. Human*. 34, 82–99. doi: 10.1093/llc/fqy017

Liao, X., and Cheng, G. (2016). *Analysing the Semantic Change Based on Word Embedding. NLPCC/ICCPOL*.

Lloyd, A. J. (2012). "The Englishmen here are much disliked': hostility towards English immigrants in early twentieth-century Toronto," in *Locating the English Diaspora*, eds T. Bueltmann, D. T. Gleeson, and D. MacRaild (Liverppol: Liverpool University Press), 135–149. doi: 10.5949/UPO9781846317712.009

Locke, T. (2004). *Critical Discourse Analysis*. London: Continuum.

Lucassen, J., Lucassen, L., and Manning, P. (Eds.). (2010). "Migration history in world history: multidisciplinary approaches," in *Studies in Global Social History* (Leiden: Brill), 3.

Macaulay, R. K. S. (1977). *Language, Social Class and Education: A Glasgow Study*. Edinburgh.

Macdonald, I. A., and Toal, R. (eds.). (2014). *Immigration Law and Practice in the United Kingdom, 9th Edn*. London: LexisNexis.

Margolis, E., and Laurence, S. (2005). "Concepts," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Stanford University). Available online at: https://plato.stanford.edu/archives/spr2014/entries/concepts/#Aca

Martinez-Ortiz, C., Kenter, T., Wevers, M., Huijnen, P., Verheul, J., and van Eijnatten, J. (2016). "Design and implementation of ShiCo: visualising shifting concepts over time," in *Proceedings of the 3th Histoinformatics Conference,* eds M. During, A. Jatowt, A. van den Bosch, and J. Preiser-Kappeller (Krakow)

Messer, M., Rene, S., and Wodak, R. (Eds.). (2012). *Migrations: Interdisciplinary Perspectives.* Wien: Springer.

Migration Observatory at the University of Oxford (2016). *The Perils of Perception and the EU*.

Mihalcea, R., and Nastase, V. (2012). "Word epoch disambiguation: finding how words change over time," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Vol. 2: Short Papers (Association for Computational Linguistics), 259–263. Available online at: https://www.aclweb.org/anthology/W11-2508

Mitra, S., Mitra, R., Suman Maity, K., Riedl, M., Biemann, C., Goyal, P., et al. (2015). An automatic approach to identify word sense changes in text media across timescales. *Nat. Lang. Eng.* 21, 773–798. doi: 10.1017/S135132491500011X

Moses, J. W. (2006). *International Migration: Globalization's Last Frontier*. Global Issues. Bangkok; Bangalore; Kuala Lumpur; Cape Town; London; New York, NY: Black Point, Nova Scotia, White Lotus, Fernwood Publishing, Books for Change, SIRD, David Philip, Zed Books, Distributed in the USA exclusively by Palgrave Macmillan.

Musolff, A. (2015). Dehumanizing metaphors in UK immigrant debates in press and online media. *J. Lang. Aggress. Conflict* 3, 41–56. doi: 10.1075/jlac.3.1.02mus

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. *arXiv [Preprint] arXiv:1802.05365.*

Pisarevskaya, A., Levy, N., Scholten, P., and Jansen, J. (2019). Mapping migration studies: an empirical analysis of the coming of age of a research field. *Migr. Stud.* mnz031. doi: 10.1093/migration/mnz031

Pocock, J. G. A. (2016). *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton, NJ: Princeton University Press.

Pooley, C. G., and Turnbull, J. (1998). *Migration and Mobility in Britain Since the Eighteenth Century*. London: UCL Press. Available online at: https://b-ok.cc/book/697408/72f250

Popescu, O., and Strapparava, C. (2014). Time corpora: epochs, opinions and changes. *Knowl. Based Syst.* 69, 3–13. doi: 10.1016/j.knosys.2014.04.029

Ramdin, R. (1999). *Reimaging Britain: Five Hundred Years of Black and Asian History*. London; Sterling, VA: Pluto Press. Available online at: http://site.ebrary.com/id/10480007

Ramdin, R. (2017). *Making of the Black Working Class in Britain.* Verso. Available online at: http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5177461

Recchia, G., Ewan, J., Nulty, P., Regan, J., and de Bolla, P. (2017). "Tracing shifting conceptual vocabularies through time," in *Knowledge Engineering and Knowledge Management*, Vol. 10180, eds P. Ciancarini, F. Poggi, M. Horridge, J. Zhao, T. Groza, M. C. Suarez-Figueroa, M. d'Aquin, and V. Presutti (Cham: Springer International Publishing), 19–28. doi: 10.1007/978-3-319-58694-6_2

Reisigl, M., and Wodak, R. (2001). "The discourse-historical approach," in *Methods of Critical Discourse Analysis*, eds R. Wodak and M. Meyer (London: Sage), 63–94.

Reisigl, M., and Wodak, R. (2009). "The discourse-historical approach," in *Methods of Critical Discourse Analysis*, 2nd Edn, eds R. Wodak and M. Meyer (London: Sage), 87–121.

Rich, P. B. (1986). *Race and Empire in British Politics. Comparative Ethnic and Race Relations*. Cambridge; New York, NY: Cambridge University Press.

Richards, E. (1993). How did poor people emigrate from the British Isles to Australia in the nineteenth century?' *J. Br. Stud.* 32, 250–279. doi: 10.1086/386032

Sharry, F. (2000). NGOs and the future of the migration debate. *J. Int. Migrat. Integr.* 1, 121–130. doi: 10.1007/s12134-000-1011-7

Skinner, Q. (1969). Meaning and understanding in the history of ideas. *History Theory* 8:3. doi: 10.2307/2504188

Skinner, Q. (1978). *The Foundations of Modern Political Thought*. Cambridge; New York, NY: Cambridge University Press.

Skinner, Q. (2012). *Liberty Before Liberalism*. Cambridge: Cambridge University Press.

Snow, C. P. (2013). *The Two Cultures and the Scientific Revolution*. Mansfield Center, CT: Martino Publishing.

Solomos, J. (2003). *Race and Racism in Contemporary Britain*. Basingstoke; Hampshire: Macmillan Education. Available online at: http://catalog.hathitrust.org/api/volumes/oclc/21950867.html

Spencer, I. R. G. (1997). *British Immigration Policy Since 1939: The Making of Multi-Racial Britain*. New York, NY: Routledge.

Spencer, S. (1994). *Strangers and Citizens: A Positive Approach to Migrants and Refugees*. London: Institute for public policy research; Rivers Oram Press.

Spencer, S. (2011). *The Migration Debate.* Bristol: Policy Press.

Sword, K. R. (1986). "Their prospects will not be bright": British responses to the problem of the polish "recalcitrants" 1946-49. *J. Contemp. Hist.* 21, 367–390. doi: 10.1177/002200948602100302

Tabili, L. (1994). The construction of racial difference in twentieth-century Britain: the special restriction (Coloured Alien Seamen) order, 1925. *J. Br. Stud.* 33, 54–98. doi: 10.1086/386044

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. *arXiv:1811.06278 Cs*.

Threadgold, T. (2009). *The Media and Migration in the United Kingdom, 1990-2009*. Washington., DC: Migration Policy Institute.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intellig. Res.* 37, 141–188. doi: 10.1613/jair.2934

van Dijk, T. (ed.). (1997). *Discourse Studies: A Multidisciplinary Introduction*, Vol. 2. London: Sage.

van Eijnatten, J., Pieters, T., and Verheul, J. (2014). Big data for global history: the transformative promise of digital humanities. *Low Count. Historic. Rev.* 128, 55–77. doi: 10.18352/bmgn-lchr.9350

Viola, L., and Musolff, A. (2019). *Migration and Media: Discourses about Identities in Crisis*. Amsteram: John Benjamins.

Viola, L., and Verheul, J. (2019). Mining ethnicity: discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. *Dig. Schol. Human.* fqz068. doi: 10.1093/llc/fqz068

Wenker, G. (1878). *Sprach-Atlas Der Rheinprovinz N?rdlich Der Mosel Sowie Des Kreises Siegen*. [Language-Atlas of the Rhine Province North of the Mosel and of the District of Siegen]. Available online at: https://scholar.google.com/scholar_lookup?hl=en&publication_year=1878&author=G.+Wenker&title=Sprach-Atlas+der+Rheinprovinz+n%C3%B6rdlich+der+Mosel+sowie+des+Kreises+Siegen

Wodak, R., and Fairclough, N. (1997). "Critical discourse analysis," in *Discourse as Social Interaction*, ed T. A. van Dijk (London: Sage), 258–284.

Wodak, R., and Meyer, M. (eds.). (2001). *Methods of Critical Discourse Analysis*. London: SAGE.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/frai.2020.00073

Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach

Yuri Bizzoni[1]*, *Stefania Degaetano-Ortlieb*[1], *Peter Fankhauser*[2] and *Elke Teich*[1]

[1] Language Science and Technology, Saarland University, Saarbrücken, Germany, [2] Digital Linguistics, Institut für Deutsche Sprache, Mannheim, Germany

We trace the evolution of Scientific English through the Late Modern period to modern time on the basis of a comprehensive corpus composed of the Transactions and Proceedings of the Royal Society of London, the first and longest-running English scientific journal established in 1665. Specifically, we explore the linguistic imprints of specialization and diversification in the science domain which accumulate in the formation of "scientific language" and field-specific sublanguages/registers (chemistry, biology etc.). We pursue an exploratory, data-driven approach using state-of-the-art computational language models and combine them with selected information-theoretic measures (entropy, relative entropy) for comparing models along relevant dimensions of variation (time, register). Focusing on selected linguistic variables (lexis, grammar), we show how we deploy computational language models for capturing linguistic variation and change and discuss benefits and limitations.

Keywords: linguistic change, diachronic variation in language use, register variation, evolution of Scientific English, computational language models

OPEN ACCESS

Edited by:
David Jurgens,
University of Michigan, United States

Reviewed by:
Joshua Waxman,
Yeshiva University, United States
Yao-Ying Lai,
National Chengchi University, Taiwan

****Correspondence:***
Yuri Bizzoni
yuri.bizzoni@uni-saarland.de

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *03 February 2020*
Accepted: *07 August 2020*
Published: *16 September 2020*

Citation:
Bizzoni Y, Degaetano-Ortlieb S,
Fankhauser P and Teich E (2020)
Linguistic Variation and Change in 250
Years of English Scientific Writing: A
Data-Driven Approach.
Front. Artif. Intell. 3:73.
doi: 10.3389/frai.2020.00073

1. INTRODUCTION

The language of science is a socio-culturally firmly established domain of discourse that emerged in the Early Modern period (ca. 1500–1700) and fully developed in the Late Modern period (ca. 1700–1900). While considered fairly stable linguistically (cf. Görlach, 2001; Leech et al., 2009), the Late Modern period is a very prolific time when it comes to the formation of text types, with many of the registers we know today developing during that period—including the language of science (see Görlach, 2004 for a diachronic overview).

Socio-culturally, register diversification is connected to the growing complexity of modern societies, labor becoming increasingly divided with more different and increasingly specialized activities across all societal sectors[1]. Also, driven by science as well as early industry, standardization (e.g., agreements on weights and measures) and routinization of procedures become important issues. At the same time, enlightenment and the scientific and industrial revolutions support a general climate of openness and belief in technological advancement. In the domain of science, the eighteenth century is of course the epoch of encyclopedias[2] but also that of the scientific academies which promoted the scientific method and distributed scientific knowledge through

[1] An example in point are production and experimentation, which used to be carried out hand in hand in the workshops of alchemists and apothecaries but were separated later on, also physically, with experimentation becoming a scientific activity carried out in dedicated laboratories (Burke, 2004; Schmidgen, 2011).

[2] For example, the publication of the famous *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* (1751–1765).

their publications. The two oldest scientific journals are the French *Journal des Sçavans* and the *Philosophical Transactions of the Royal Society of London*. At the beginning of publication (both started in 1665), the journals were no more than pamphlets and included articles written in the form of letters to the editor and reviews of scientific works (Gleick, 2010). Professionalization set in around the mid eighteenth century, as witnessed by the introduction of a reviewing process in the Royal Society (Moxham and Fyfe, 2018; Fyfe et al., 2019).

While there is a fair stock of knowledge on the development of scientific language from socio-cultural and historical-pragmatic perspectives (see section 2), it is less obvious what are the underlying, more general principles of linguistic adaptation to new needs of expression in an increasingly diversified and specialized setting such as science. This provides the motivation for the present research. Using a comprehensive diachronic corpus of English scientific writing composed of the Philosophical Transactions and Proceedings of the Royal Society of London [henceforth: Royal Society Corpus (RSC); Kermes et al., 2016; Fischer et al., 2020], we trace the evolution of Scientific English looking for systematic linguistic reflexes of specialization and diversification, yielding a distinctive "scientific style" and forming diverse sublanguages (sublanguage of chemistry, physics, biology etc.). In terms of theory, our work is specifically rooted in register linguistics (Halliday, 1985b; Biber, 1988) and more broadly in theories of language use, variation and change that acknowledge the interplay of social, cognitive and formal factors (e.g., Bybee, 2007; Kirby et al., 2015; Aitchison, 2017; Hundt et al., 2017). While we zoom in on the language of science, we are ultimately driven by the more general questions about language change: What changes and how? What drives change? How does change proceed? What are the effects of change? Thus, we aim at general insights about the dynamics of language use, variation and change.

In a similar vein, the methodology we present can be applied to other domains and related analysis tasks as well as other languages. Overall, we pursue an exploratory, data-driven approach using state-of-the-art computational language models (ngram models, topic models, word embeddings) combined with selected information-theoretic measures (entropy, relative entropy) to compare models/corpora along relevant dimensions of variation (here: time and register) and to interpret the results with regard to effects on language system and use. Since the computational models we use are word-based, words act as the anchor unit of analysis. However, style is primarily indicated by lexico-grammatical usage, so we investigate both the lexical and the grammatical side of words. While we consider lexis and grammar as intricately interwoven, in line with various theories of grammar (Halliday, 1985a; Hunston and Francis, 2000; Goldberg, 2006), for expository purposes, we here consider the lexico-semantic and the lexico-grammatical contributions to change separately.

The remainder of the paper is organized as follows. We start with an overview of previous work in corpus and computational linguistics in modeling diachronic change with special regard to register and style (section 2). In section 3 we introduce our data set (section 3.1) and elaborate on the methods employed (section 3.3). Section 4 shows analyses of diachronic trends at the levels of lexis and grammar (section 4.1), the development of topics over time (section 4.2) and paradigmatic effects of changing language use (section 4.3). Finally, we summarize our main results and briefly assess benefits and shortcomings of the different kinds of models and measures applied to the analysis of linguistic variation and change (section 5).

2. RELATED WORK

The present work is placed in the area of language variation and change with special regard of social and register variation and computational models of variation and change (for overviews see Aragamon, 2019 for computational register studies and Nguyen et al., 2016 for computational socio-linguistics).

Regarding the language of science, there is an abundance of linguistic-descriptive work, including diachronic aspects, providing many valuable insights (e.g., Halliday, 1988; Halliday and Martin, 1993; Atkinson, 1999; Banks, 2008; Biber and Gray, 2011, 2016). However, most of the existing work is either based on text samples or starts from predefined linguistic features. Further, there are numerous studies on selected scientific domains, such as medicine or astronomy, e.g., Nevalainen (2006), Moskowich and Crespo (2012) and Taavitsainen and Hiltunen (2019), which work on the basis of fairly small corpora containing hand-selected and often manually annotated material. Typically, such studies are driven from a historical socio-linguistic or pragmatic perspective and focus on selected linguistic phenomena, e.g., forms of address (Taavitsainen and Jucker, 2003). For overviews on recent trends in historical pragmatics/socio-linguistics (see Jucker and Taavitsainen, 2013; Säily et al., 2017). Studies on specific domains, registers or text types provide valuable resources and insights into the socio-historical conditions of language use. Here, we build upon these insights, adding to it the perspective of general mechanisms of variation and change.

More recently, the diachronic perspective has attracted increasing attention in computational linguistics and related fields. Generally, diachronic analysis requires a methodology for comparison of linguistic productions along the time line. Such comparisons may range over whole epochs (e.g., systemic changes from early Modern English to Late Modern English), or involve short ranges (e.g., the issues of 1 year of The New York Times to detect topical trends). Applying computational language models to diachronic analysis requires a computationally valid method of comparison of language use along the time line, i.e., one that captures linguistic change if it occurs.

Different kinds of language models are suitable for this task and three major strands can be identified. First, a number of authors from fields as diverse as literary studies, history and linguistics have used simple ngram models to find trends in diachronic data using relative entropy (Kullback-Leibler Divergence, Jensen-Shannon Divergence) as a measure of comparison. For instance, Juola (2003) used Kullback-Leibler Divergence (short: KLD) to measure rate of linguistic change in

30 years of National Geographic Magazine. In more recent, large-scale analyses on the Google Ngram Corpus (Bochkarev et al., 2014; Kim et al., 2014) analyze change in frequency distributions of words within and across languages. Specifically humanistic research questions are addressed by e.g., Hughes et al. (2012) who use relative entropy to measure stylistic influence in the evolution of literature; or Klingenstein et al. (2014) who analyze different speaking styles in criminal trials comparing violent with non-violent offenses; or Degaetano-Ortlieb and Teich (2018) applying KLD as dynamic slider over the time line of a diachronic corpus of scientific text.

Second, probabilistic topic models (Steyvers and Griffiths, 2007) have become a popular means to summarize and analyze the content of text corpora, including topic shifts over time. In linguistics and the digital humanities, topic models have been applied to various analytic goals including diachronic linguistic analysis (Blei and Lafferty, 2006; Hall et al., 2008; Yang et al., 2011; McFarland et al., 2013). Here again, a valid method of comparing model outputs along the time line has to be provided. In our work, we follow the approach proposed in Fankhauser et al. (2016) using entropy over topics as a measure to assess topical diversification over time.

Third, word embeddings have become a popular method for modeling linguistic change, with a focus on lexical semantic change (e.g., Hamilton et al., 2016; Dubossarsky et al., 2017, 2019; Fankhauser and Kupietz, 2017). Word embeddings are weakly neural models that capture usage patterns of words and are used in a variety of NLP tasks. While well-suited to capture the summative effects of change (groups of words or whole vocabularies, see e.g., Grieve et al., 2016), the primary focus lies on lexis[3]. Other linguistic levels, e.g., grammar (Degaetano-Ortlieb and Teich, 2016, 2018; Bizzoni et al., 2019a), collocations (Xu and Kemp, 2015; Garcia and Garćia-Salido, 2019), or specific aspects of change, e.g., spread of change (Eisenstein et al., 2014), specialization (Bizzoni et al., 2019b) or life-cycles of language varieties (Danescu-Niculescu-Mizil et al., 2013), are only rarely considered. Once again, while word embeddings offer a specific model of language use, using them to capture diachronic change and to assess effects of change calls for adequate instruments for comparison along the time line. Here, we use the commonly applied measure of cosine distance for a general topological analysis of diachronic word embedding spaces; and we use entropy for closer inspection of specific word embeddings clusters to measure the more fine-grained paradigmatic effects of change.

In sum, in this paper we address some of the core challenges in modeling diachronic change by (a) looking at the *interplay* of different linguistic levels (here: lexis and grammar), (b) elaborating on the formation of style and register from a diachronic perspective, and (c) enhancing existing computational methods with explicit measures of linguistic change. Since we are driven by the goal of explanation rather than high-accuracy prediction (as in NLP tasks), qualitative interpretation by humans is an integral step. Here, micro-analytic and visual support are doubly important if one wants to explore linguistic conditions and effects of change. To support this, good instruments for human inspection and analysis of data are crucial—see, for instance, Jurish (2018) and Kaiser et al. (2019) providing visualization tools for various aspects of diachronic change, partly with interactive function (Fankhauser et al., 2014; Fankhauser and Kupietz, 2017); or Hilpert and Perek (2015)'s application of motion charts to the analysis of meaning change. We developed a number of such visualization tools made available as web applications for inspection of the Royal Society Corpus (cf. section 3).

TABLE 1 | Size of RSC 6.0 by 50-year periods.

Time	# Tokens	# Texts
1665–1699	2,582,856	1,325
1700–1749	3,414,795	1,686
1750–1799	6,342,489	1,819
1800–1849	9,112,274	2,774
1850–1899	36,993,412	6,754
1900–1919	19,273,112	3,049

3. DATA AND METHODS

3.1. Data

The corpus used for the present analysis is the Royal Society Corpus 6.0 (Fischer et al., 2020). The full version is composed of the Philosophical Transactions and Proceedings of the Royal Society from 1665 to 1996. In total, it contains 295,895,749 tokens and 47,837 documents. Here, we use a version that is open-source under a creative commons license covering the period of 1665 to 1920. In terms of periods of English, this reflects the Late Modern period (1700–1900) plus a bit of material from the last decades of the Early Modern period (before 1700) as well as a number of documents from modern English. Altogether this open version contains 78,605,737 tokens and 17,520 documents.

Note that the RSC is not balanced, later periods containing substantially more material than earlier ones (see **Table 1**), which calls for caution regarding frequency effects. Other potentially interesting features of the corpus are that the number of different authors increases over time; so does the number of papers with more than one author.

The documents in the corpus are marked up with meta-data including author, year of publication, text type and time period (1-, 10-, 50-year periods). The corpus is tokenized, lemmatized, annotated with part-of-speech tags and normalized (keeping both normalized and original word forms) using standard tools (Schmid, 1995; Baron and Rayson, 2008). The corpus is made available under a Creative Commons license, downloadable and

[3] For more comprehensive overviews on computational approaches to lexical semantic change see Tahmasebi et al. (2018) and on diachronic word embeddings see Kutuzov et al. (2018).

accessible via a web concordance (CQPWeb; Hardie, 2012) as well as interactive visualization tools[4].

3.2. Methods

There are two important a priori considerations regarding modeling linguistic change and variation. First, one of the key concepts in language variation is *use in context*. Apart from extra-linguistic, situational context (e.g., field, tenor, and mode; Quirk et al., 1985), intra-linguistic context directly impacts on linguistic choice, both syntagmatically (as e.g., in collocations) and paradigmatically (i.e., shared context of alternative expressions). Different computational models take into account different types of context and accordingly reveal different kinds of linguistic patterns. Topic models take into account the distribution of words in document context and are suitable to capture the field of discourse (see section 3.2.2 below). Plain ngram models take into account the immediately preceding words of a given word and can reveal syntagmatic usage patterns (see section 3.2.1 below). Word embeddings take into account left and right context (e.g., ± five words) and allow clustering words together depending on similar, surrounding contexts; thus, they are suited for capturing linguistic paradigms (see section 3.2.3 below).

Second, diachronic linguistic analysis essentially consists of *comparison of corpora* representing language use at different time periods. Computational language models being representations of corpora, the core task consists in comparing model outputs and elicit significant differences between them. Common measures of comparing language models are perplexity and relative entropy, typically used for assessing the quality or fit of a model by estimating the difference between models in bits using a log base. Here, we use the asymmetric version of relative entropy, Kullback-Leibler Divergence, to assess differences between language models according to time. An intimately related measure is entropy. Entropy considers the richness and (un)evenness of a sample and is a common means to measure diversity, e.g., the lexical diversity of a language sample (Thoiron, 1986). Here, we use entropy as a measure of diversification at two levels, the level of topics (field of discourse) and the level of paradigmatic word clusters, where greater entropy over time is interpreted as a signal of linguistic diversification and lower entropy as a signal of consolidated language use. The most basic way of exploring change in a given data set is to test whether the entropy over a simple bag-of-words model changes or not. For diversification to hold, we would expect the entropy to rise over time in the RSC, also because of the increase in size of the more recent corpus parts as well as in number of authors. As will be seen, this is not the case, entropy at this level being fairly stable (section 4.2).

3.2.1. Ngram Based Models

To obtain a more fine-grained and linguistically informed overview of the overall diachronic tendencies in the RSC than possible with token ngrams, we consider lexical and grammatical usage separately using lemmas and part-of-speech (POS) sequences as modeling units. On this basis, models of different time periods (e.g., decades) are compared with the asymmetric variant of relative entropy, Kullback-Leibler Divergence (KLD; Kullback and Leibler, 1951); cf. Equation (1) where A and B here denote different time periods.

$$D(A||B) = \sum_i p(unit_i|A) log_2 \frac{p(unit_i|A)}{p(unit_i|B)} \tag{1}$$

KLD is a common measure for comparing probability distributions in terms of the number of additional bits needed for encoding when a non-optimal model is used. Applied to diachronic comparison, we obtain a reliable index of difference between two corpora A and B: the higher the amount of bits, the greater the diachronic difference. Also, we know which specific units/features contribute to the overall KLD score by their pointwise KLD. Thus, we can inspect particular points in time (e.g., by ranking features by pointwise KLD in 1 year) or time spans (e.g., by standard deviation across several years) to dynamically observe changes in a feature's contribution. This gives us two advantages over traditional corpus-based approaches: no predefined features are needed and results are more directly interpretable.

Apart from comparing predefined time periods with each other as is commonly done in diachronic corpus-linguistic studies (cf. Nevalainen and Traugott, 2012 for discussion), KLD can be used as a data-driven periodization technique (Degaetano-Ortlieb and Teich, 2018, 2019). KLD is dynamically pushed over the time line comparing past and future (or, as KLD is asymmetric, future vs. past). As we will show below, using KLD in this way allows detecting diachronic trends that are hard to see on a token level or with predefined, more coarse time periods. The granularity of diachronic comparison can be varied depending on the corpus and the analytic goal (year-, month-, day-based productions); again, no a priori assumptions have to be made regarding the concrete linguistic features involved in change other than selecting the linguistic level of comparison (e.g., lemmas, parts of speech). Hence, the method is generic and at the same time sensitive to the data.

3.2.2. Topic Models

To obtain a picture of the diachronic development in terms of field of discourse—a crucial component in register formation—we need to consider the usage of words in the context of whole documents. To this end, we use topic models. We follow the overall approach of applying topic models to diachronic corpora mapping topics to documents (Blei and Lafferty, 2006; Steyvers and Griffiths, 2007; Hall et al., 2008; Yang et al., 2011; McFarland et al., 2013). The principle idea is to model the generation of documents with a randomized two-stage process: For every word w_i in a document d select a topic z_k from the document-topic distribution $P(z_k|d)$ and then select the word from the topic-word distribution $P(w_i|z_k)$. Consequently, the document-word distribution is factored as: $P(w_i|d) = \sum_k P(w_i|z_k)P(z_k|d)$. This factorization effectively reduces the dimensionality of the model for documents, improving their interpretability: Whereas $P(w_i|d)$ requires one dimension for each distinct word (tens of thousands) per document, $P(z_k|d)$

[4] RSC 6.0 Open: https://hdl.handle.net/21.11119/0000-0004-8E37-F.

only requires one dimension for each topic (typically in the range of 20–100). To estimate the document-topic and topic-word distributions from the observable document-word distributions we use Gibbs-Sampling as implemented in MALLET[5].

To investigate topical trends over time, we average the document-topic distributions for each year y:

$$P(z_k|y) = 1/n \sum_{d_j \in y} P(z_k|d_j) \tag{2}$$

where n is the number of documents per year.

For further interpretation, we cluster topics hierarchically on the basis of the distance[6] between their topic-document distributions (Equation 3).

$$P(d|z) = P(z|d) / \sum_j P(z|d_j) \tag{3}$$

Topics that typically co-occur in documents have similar topic-document distributions, and thus will be placed close in the cluster tree.

To assess diachronic diversification in discourse field as a central part of register formation, we measure the entropy over topics (cf. Equation 4), and the mean entropy of topic-word distributions per time period.

$$H(P(.|y)) = -\sum_k P(z_k|y) log_2 P(z_k|y) \tag{4}$$

Note that all measures operate on relative frequencies per time period in order to control for the lack of balance in our data set (more recent periods contain considerably more data than earlier ones).

3.2.3. Word Embeddings

Word embeddings (WEs) capture lexical paradigms, i.e., sets of words sharing similar syntagmatic contexts. Word embeddings build on the principle underlying distributional semantics that it is possible to capture important aspects of the semantics of words by modeling their context (Harris, 1954; Lenci, 2008).

Here, we apply WEs diachronically to explore the overall development of word paradigms in our corpus with special regard to register/sublanguage formation as well as scientific style. Using the approach and tools provided by Fankhauser and Kupietz (2017) we compute WEs with a structured skip-gram approach (Ling et al., 2015). This is a variant of the popular Word2Vec approach (Mikolov et al., 2013). Word2Vec is a way of maximizing the likelihood of a word given its context, by training a d x V matrix where V is the vocabulary and d an arbitrary number of dimensions.

The goal of the algorithm is to maximize

$$L = \frac{1}{T} \sum_{t \in T} \sum_{-c \leq j \leq c} log\ p(w_{t+j}|w_t) \tag{5}$$

where T is a text and c is the number of left and right context words to be taken into consideration. In short, the model tries to learn the probability of a word given its context, $p(w_o|w_i)$. To this end, the model learns a set of weights that maximizes the probability of having a word in a given context. Such set of weights constitutes a word's embedding.

Usually, skip-gram considers a term's context as a bag-of-words. In Ling et al. (2015)'s variant, the order of the word context is also taken into consideration which is important to capture words with grammatical functions rather than lexical words only. For diachronic application, we calculate WEs per time period (e.g., 1-/10-/50-year periods), where the first period is randomly initialized, and each subsequent period is initialized by the model for its preceding period. Thereby, WEs are comparable across periods.

To perform analyses on our models, we then apply simple similarity measures commonly used in distributional semantics, where the similarity between two words is assessed by the cosine similarity of their vectors:

$$sim(w_1, w_2) = cos(w_1, w_2) = \frac{w_1 w_2}{|w_1||w_2|} \tag{6}$$

where w_1 and w_2 are the vectors of the two words taken into consideration, and $|w|$ is a vector's norm. Alternatively, the semantic distance between words can be considered, which is the complement of their similarity:

$$dist(w_1, w_2) = 1 - cos(w_1, w_2) \tag{7}$$

To detect the semantic tightness or level of clustering of a group of words (how semantically similar they are), one can thus compute the average cosine similarity between all the words in a group of words:

$$sim(V) = \frac{\sum_{w_a \in V} \sum_{w_b \in V} cos(w_a, w_b)}{V^2} \tag{8}$$

where V (vocabulary) is the group of words taken into consideration. Reversely, it is possible to compute the average distance of a group of words from another group of words by iterating the sums on two different sets.

To detect semantic shifts over time, one of the simplest and most popular approaches is that of computing the change of the cosine similarity between a group of pre-defined words in a chronologically ordered set of WE spaces. As we will show, the WE space of the RSC as a whole expands over time. At the same time, it becomes more *fragmented* and specific clusters of words become more densely populated while others disappear. We base such observations on an analysis of the word embeddings' topology using cosine similarity as explained above as well as entropy. For example, since the period under investigation witnesses the systematization of several scientific disciplines, we are likely to observe a narrowing of the meaning of many individual words—mainly technical terms—which would push them further away from one another. Similarly, for specific WE clusters, we expect growth or decline, e.g., chemical terms explode in the late eighteenth century, pointing to the emergence

[5] http://mallet.cs.umass.edu

[6] We use Pearson distance, which consistently results in more intuitive hierarchies than Jensen-Shannon Divergence.

of the field of chemistry with the associated technical language, or many Latin words disappear. Such developments can be measured by the entropy $H(P(.|w))$ over a given cluster around word w, by estimating the conditional probability of words w_i in the close neighborhood of word w as follows:

$$P(w_i|w) = sim(w, w_i) * freq(w_i, w)/(\sum_k sim(w, w_k) * freq(w_k, w)) \quad (9)$$

where w_k ranges over all words (including w) with sufficient similarity (e.g., > 0.6) to w. The neighbors are weighted by their similarity to the given word, thus, a word with many near neighbors and rather uniform distribution has a large entropy, indicating a highly diversified semantic field.

4. ANALYSES

Our analyses are driven by two basic assumptions: register diversification (linguistic variation focused on field of discourse) and formation of “scientific style” (convergence on specific linguistic usages within the scientific domain). We carry out three kinds of analysis on the Royal Society Corpus showing these two major diachronic trends at the levels of lexis and grammar (section 4.1), development of topic over time (section 4.2) as well as paradigmatic effects (section 4.3).

4.1. Diachronic Trends in Lexis and Grammar

We trace the overall diachronic development in the RSC considering both lexical and grammatical levels. Lexis is captured by lemmas and grammar by sequences of three parts of speech (POS). Using the data-driven periodization technique described in section 3.2.1 based on KLD, we dynamically compare probability distributions of lemma unigrams and POS trigrams along the time line.

Figures 1A,B plot the temporal development for the lexical and the grammatical level, respectively. The black line visualizes relative entropy of the future modeled by the past, i.e., how well at a particular point in time the future can be modeled by a model of the past (here: 10 year slices). The gray line visualizes the reverse, i.e., how well the past is modeled by the future (again on 10-year slices). Peaks in the black line indicate changes in the future which are not captured by a model of the past, such as new terminology. Peaks in the gray line indicate differences from the opposite perspective, i.e., the future not encompassing the past, e.g., obsolete terminology. Troughs for both lines indicate convergence of future and past. A fairly persistent, low-level relative entropy indicates a period of stable language use.

Comparing the two graphs in **Figure 1**, we observe a particularly strong decreasing tendency for the grammatical level (see **Figure 1B**) and a slightly declining tendency at the lexical level with fairly pronounced oscillations of peaks and troughs (**Figure 1A**). Basically, peaks indicate innovative language use, troughs indicate converging use, the future being less and less “surprised” by the past. Thus, while grammatical usage consolidates over time, the lexical level is more volatile as it reacts directly to the pressure of expressing newly emerging things or concepts in peoples’ (changing) domains of experience (here: new scientific discoveries). The downward trend at the grammatical level is a clear sign of convergence, possibly related to the formation of a scientific style; peaks at the lexical level signal innovative use and may indicate register diversification.

To investigate this in more detail, we look at specific lexical and grammatical developments. We use pointwise KLD (i.e., the contribution of individual features to overall KLD) to rank features. For example, there is a major increase in overall KLD around the 1790s at the lemma level. Considering features contributing to the highest peak in 1791 for the FUTURE model (black line), we see a whole range of words from the chemistry field around *oxygen* (see **Figure 2**). At the same time, we can inspect which features leave language use and contribute to an increase in KLD for the PAST model (i.e., features not well-captured by the future anymore). From **Figure 3**, we observe words related to *phlogiston* and experiments with air contributing to the formation of the oxygen theory of combustion (represented by Lavoisier, Priestley as well as Scheele). In fact, the oxygen theory replaced Becher and Stahl’s 100-years

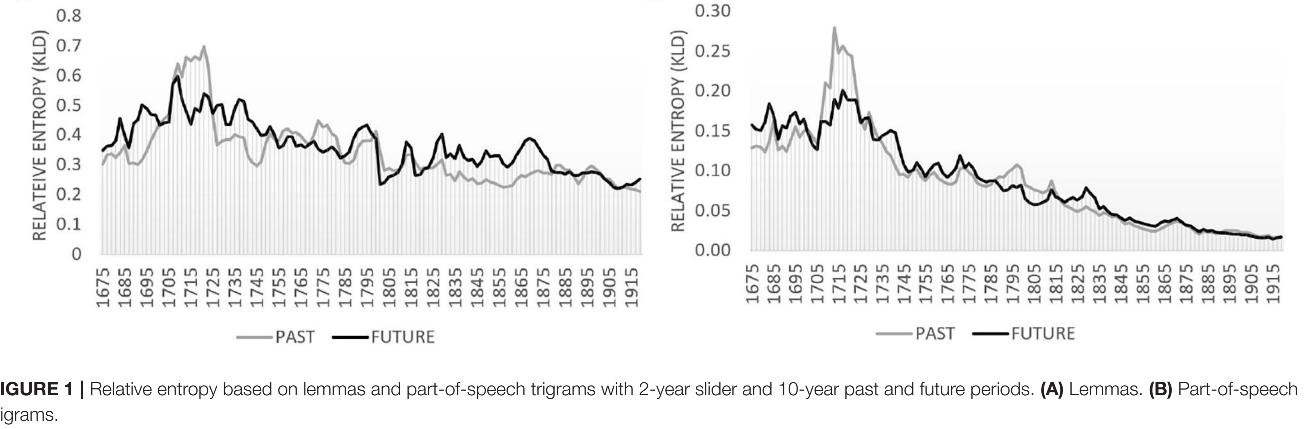


FIGURE 1 | Relative entropy based on lemmas and part-of-speech trigrams with 2-year slider and 10-year past and future periods. **(A)** Lemmas. **(B)** Part-of-speech trigrams.

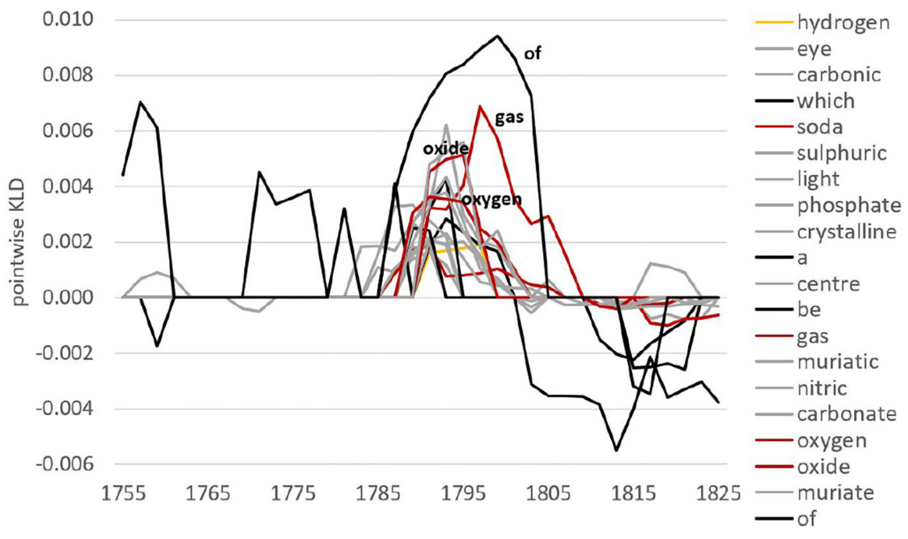


FIGURE 2 | Pointwise relative entropy based on lemmas for the FUTURE model in 1791.

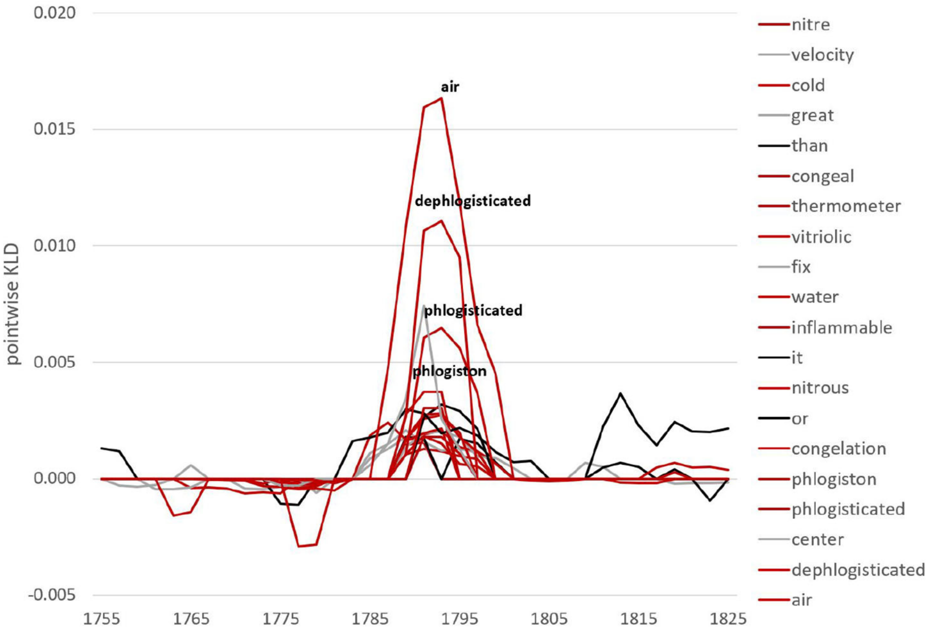


FIGURE 3 | Pointwise relative entropy based on lemmas for the PAST model in 1791.

old phlogiston theory, marking a chemical revolution in the eighteenth century—it is this shift of scientific paradigm that we encounter here in the RSC.

At the grammatical level, after a fairly high KLD peak in the early 1700's, there is a step-wise, steady decrease with only local, smaller peaks. As an example of a typical development at the grammatical level consider the features involved in the 1771 peak (see **Figure 4**). These are passive voice and relational verb patterns (e.g., NOUN-BE-PARTICIPLE as in *air is separated*; blue), nominal patterns with prepositions [e.g., indicating measurements such as the NOUN-PREPOSITION-ADJECTIVE as in *the quantity of common (air)*; gray], gerunds (e.g., NOUN-PREPOSITION-*ing*VERB, such as *method of making*; yellow), and relative clauses (e.g., DETERMINER-NOUN-RELATIVIZER, such as *the air which/that*; red). While the contribution of these patterns to the overall KLD is high in 1771, it becomes zero for all of them by 1785—a clear indication of consolidation in grammatical usage pointing to the development of a uniform scientific style.

Regarding the lexical level, to verify that the observed tendencies point to significant diversification effects, we need to

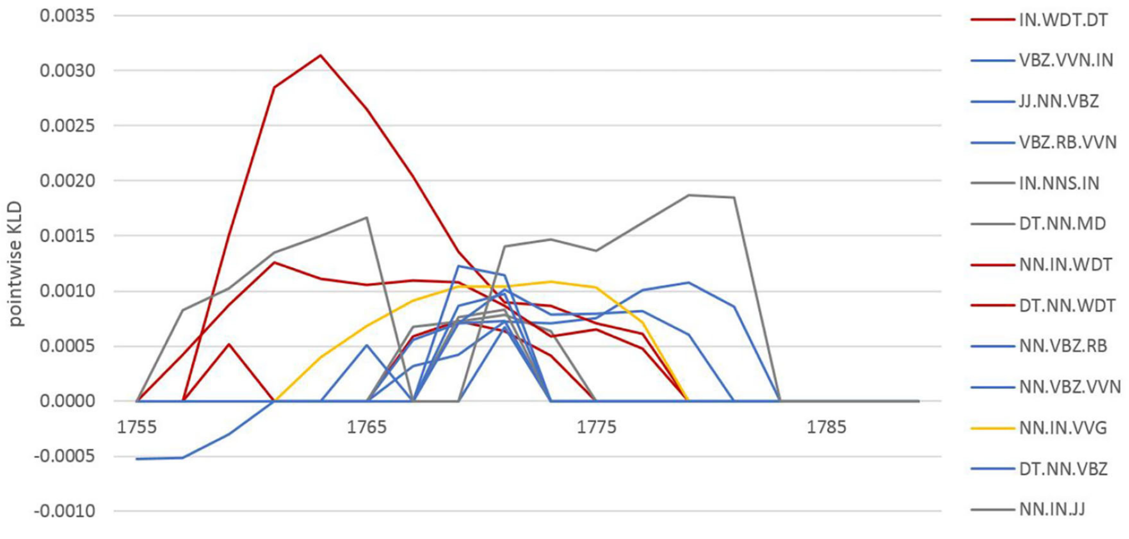


FIGURE 4 | Pointwise relative entropy based on POS trigrams for the PAST model in 1771.

explore the systematic association of words with discourse fields. For this, we turn to topic models.

4.2. Diachronic Development of Discourse Fields

To analyse the development of discourse fields over time as the core component in register diversification, we trained a topic model with 30 topics[7]. Stop words were excluded and documents were split into parts of at most 5000 tokens each to control for largely varying document lengths.

Table 2 shows four of the 30 topics with their most typical words. Note that topics do not only capture the field of discourse (BIOLOGY 3) but also genre (REPORTING), mode (FORMULAE), or simply reoccurring boiler plate text (HEADMATTER).

Figure 5A displays the topic hierarchy resulting from clustering the topics based on the Pearson Distance between their topic-document distributions[8]. Labels for topics and topic clusters have been assigned manually, and redundant topics with very similar topic word distributions, such as BIOLOGY, have been numbered through.

Figure 5B shows the probabilities of the combined topics over time. As can be seen, the first hundred years are dominated by the rather generic combined topic REPORTING, which covers around 70% of the topic space. Indeed, the underlying topic REPORTING makes for more than 50% of the topic space during the first 50 years. Starting in 1750, topics become more diversified into individual disciplines, indicating register diversification in terms of discourse field. In addition, in line with the analysis in section 3.1, we clearly see the rise of the CHEMISTRY topic around the 1790s.

TABLE 2 | Top five words for selected topics.

REPORTING	HEADMATTER	BIOLOGY 3	FORMULAE
great	vol	cells	equation
time	society	fig	equations
made	london	cell	function
found	author	tissue	form
account	part	nucleus	cos

As shown in **Figure 6A** diversification is evidenced by the clearly increasing entropy of the topic distribution over time. However, the mean entropy of the individual document-topic distributions remains remarkably stable, even though the mean number of authors per document and document length increase over time. Even the mean entropy weighted by document length (not shown) remains stable. This may be in part due to using asymmetric priors for the document-topic distributions, which generally skews them toward topics containing common words shared by many documents (Wallach et al., 2009), thus stabilizing the document-topic distributions over time.

Figure 6B shows the diachronic development of entropies at the level of words. The overall entropy of the unigram language model as well as the mean entropy of the topic word distributions weighted by the topic probabilities are also remarkably stable. However, the (unweighted) mean entropy of topic word distributions clearly increases over time. Indeed, due to the fairly high correlation of 0.81 (Spearman) between topic probability and the topic word entropy, evolving topics with increasing probability also increase in their word entropy, i.e., their vocabulary becomes more diverse. **Figure 7** demonstrates this for the evolving topics in the group LIFESCIENCE 2. All topics increase over time both

[7] For the corpus at hand, a smaller number of topics leads to conflated topics, a larger number to redundant topics.

[8] Clustering by Jensen-Shannon Divergence results in a less intuitive hierarchy.

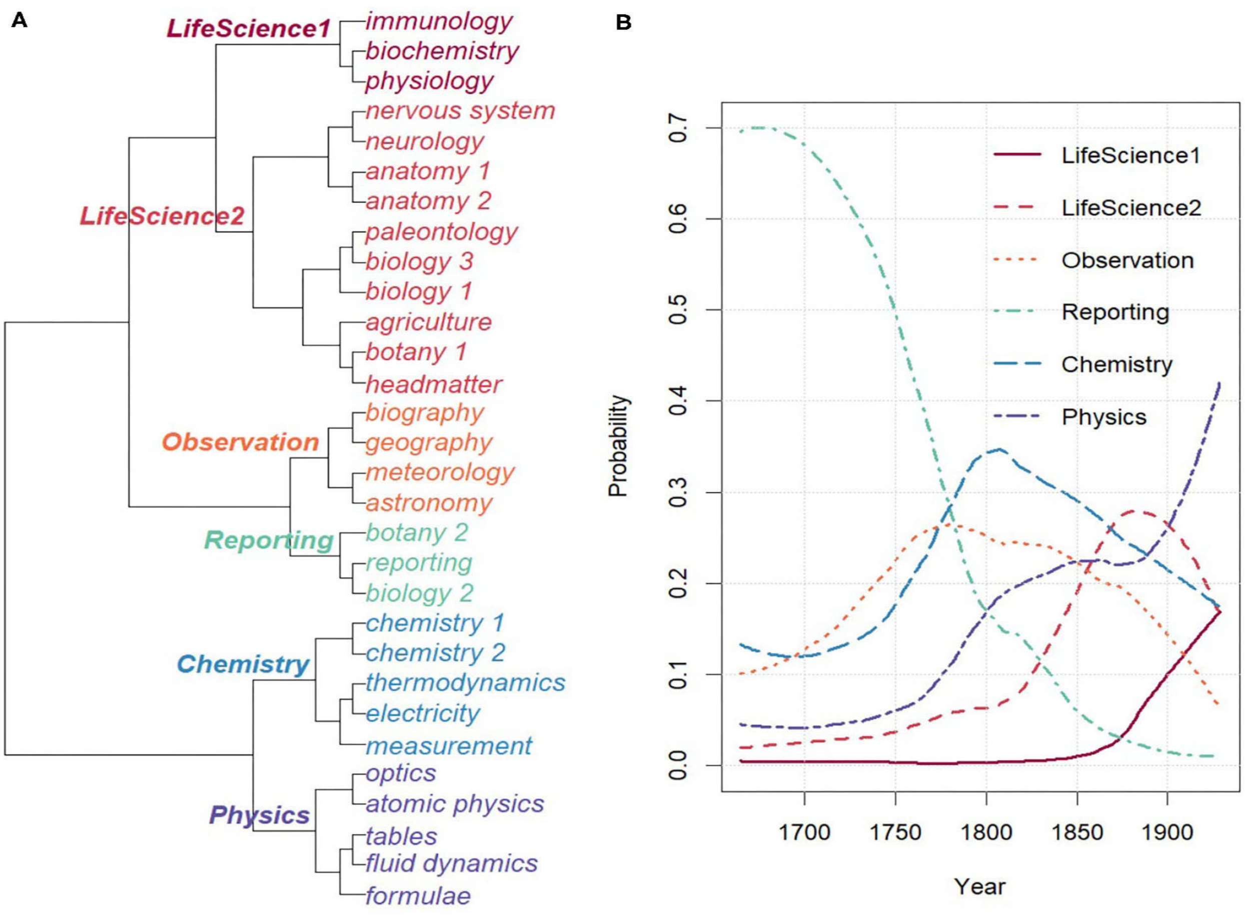


FIGURE 5 | Overview on topics. **(A)** Topic hierarchy. **(B)** Combined topics over time.

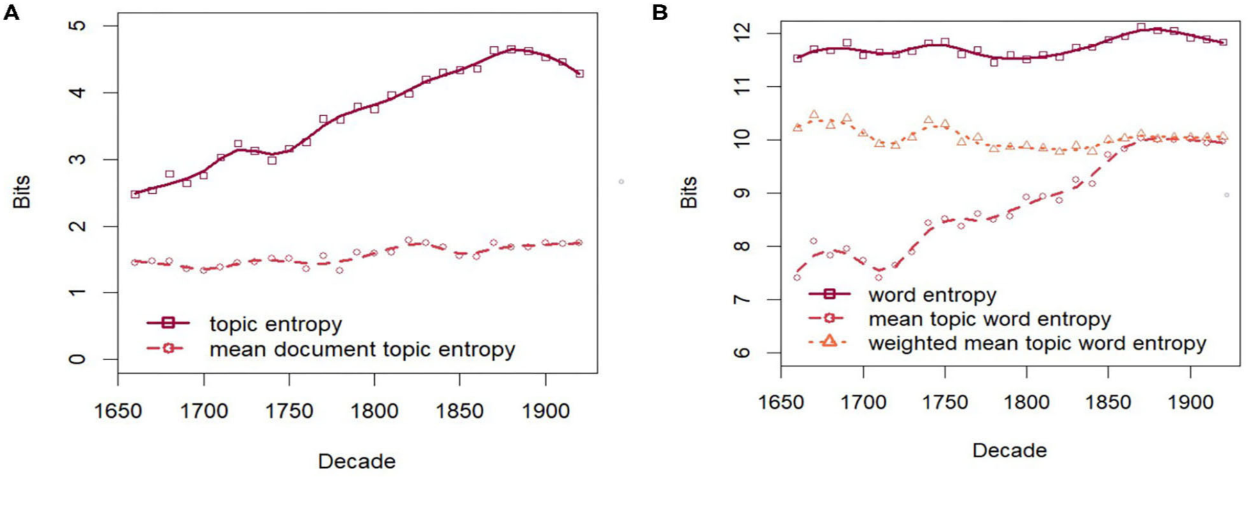


FIGURE 6 | Entropies over time. **(A)** Entropy of topics. **(B)** Entropy of words.

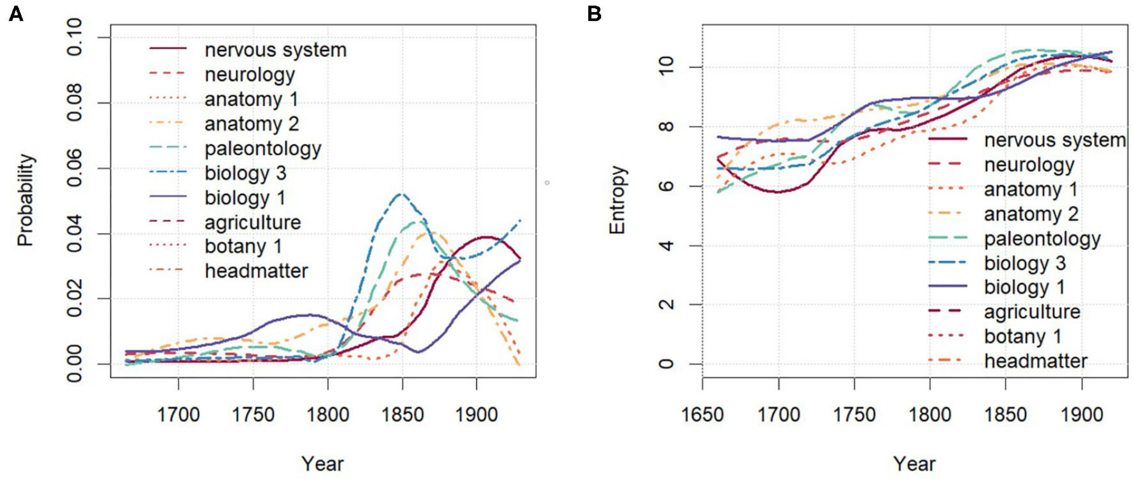


FIGURE 7 | LIFESCIENCE 2 over time. **(A)** Probability. **(B)** Entropy of topic word distributions.

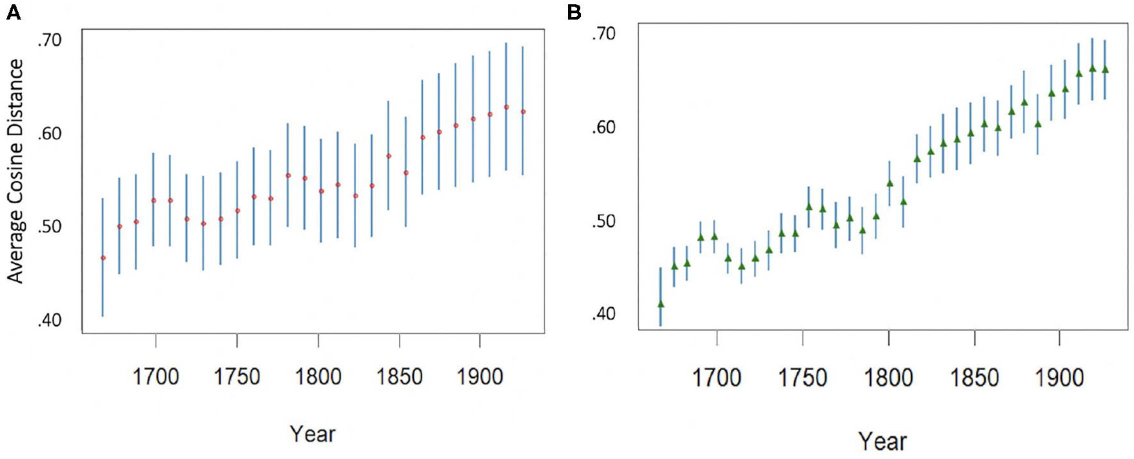


FIGURE 8 | (A) Average distance and standard deviation of 2,000 randomly selected pairs of words. **(B)** Average distance from the whole vocabulary (mean and standard deviation) of 1,000 randomly selected words.

in probability and entropy[9]. As will be seen in section 4.3, this trend is mirrored in the analysis of paradigmatic word clusters by word embeddings.

4.3. Paradigmatic Effects

To gain insights into the paradigmatic effects of the diachronic trends detected by the preceding analyses, we need to consider word usage according to syntagmatic context. To capture grammatical aspects as well (rather than just lexical-semantic patterns), we take word forms rather than lemmas as a unit for modeling and we do not exclude function words.

Based on the word embedding model as shown in section 3.2.3, we observe that the word embedding space of the RSC grows over time both in terms of *vocabulary size* and in terms of *average distance* between words. While a growing vocabulary can be interpreted in many ways, it is more informative to look at the increase in average distance between words. Here, not every term grows apart from all other terms (in fact, many pairs of words get closer through time) but when we take two random terms the average distance between them is likely to increase—see **Figure 8**: (A) shows the diachronic trend for the distance between 2,000 randomly selected pairs of words and (B) for the distance of 1,000 randomly selected words from the rest of the vocabulary. The words were selected among those terms that appear at least once in every decade. In both cases, the trend toward a growing distance is clearly visible.

Given that WEs are based on similarity in context, this means that overall, words are used increasingly in different contexts, a clear sign of diversification in language use. For example, the

[9] A similar correlation between probability and entropy can be observed in other rising topic groups.

usage of *magnify* and *glorify* diverges through the last centuries resulting in a meaning shift for *magnify* which becomes more associated with the aggrandizing effects of optical lenses while *glorify* remains closer to its original sense of elevating or making glorious. If we look for these two words in the WE space, what we see is, in fact, a progressive decrease of the distributional similarity between them: for example, in 1860 their cosine distance is 0.48, while in 1950 it has gone up to 0.62. The nature of their nearest neighbors also diverges: *magnify* increasingly shows specialized, optic-related neighborhoods (*blood-globule* in 1730, *object-lens* in 1780, *eyeglass* in 1810) while the neighbors of *glorify* remain more mixed (mainly specific but non-technical verbs, such as *bill*, *reread*, *ingratiate*, with low similarity). Finally, their movement with respect to originally close neighbors is also consistent: e.g., the distance between *glorify* and *exalt* does not change between 1860 and 1920, while *magnify* appears to move away and back toward *exalt* through the decades and is more than 25 degrees further from it in 1920 than in 1670 (from 0.45 to 0.70).

To provide another example, a similar evolution is apparent for *filling* and *saturating*: their distance grows from 0.37 in 1700 to 0.65 in 1920, a difference of almost 30 degrees. In the same lapse of time, the distance between *saturating* and *packing* goes from 0.27 to 0.70. Actually, the meaning of *saturating* was originally closer to that of *satisfying* and *packing*: its usage as a synonym of *imbuing*, and its technical sense in chemistry are more recent, and have progressively drawn the word's usage apart from that of *filling*.

As noted above, we observe an overall expansion of the WE space. To test whether this expansion is not a simple effect of the increase of frequency and number of words in each decade, we select a set of function words which exhibit stable frequency and should not change in usage over time (e.g., the functions of *the*, *and*, and *for* did not change in the period considered). If the expansion we observe is due to raw frequency effects, function words should drift apart from each other at a similar rate as content words. This appears not to be the case. As shown in **Table 3**, if we compare the group of function words to a group of randomly selected content words, such as verbs and nouns, we can see that the distances between the elements of such group grow much faster than the distances between function words. Purely functional words drift apart considerably less than words having a lexical meaning, indicating that the latter are probably causing most of the lexical expansion. Thus, words having a proper lexical meaning grow apart much faster *on average* than words having a purely functional role.

This behavior is not consistent with a raw frequency effect, or with the side effects of changes in the magnitude of training data. It looks like the distributional profile of words is, on average, growing more distinct in this specific corpus. And this does not happen only for new vocabulary, created *ad hoc* for specific contexts: even when we factor out the changes in lexicon and we consider only those words that appear in every decade (Persistent Vocabulary in **Table 3**), the effect is still visible. This interpretation is supported when we inspect the entropy on specific WE clusters over time. We consider two cases: increasing and decreasing entropy on a cluster, the former signaling lexical diversification, the latter signaling converging linguistic usage. For instance, coming back to the field of chemistry, we observe increasing entropy in particular clusters of content words: see **Figure 9** for an example, showing (A) relative frequency of selected terms denoting chemical compounds and (B) entropy on the WE cluster containing those terms (radius of cosine similarity > 0.6).

As an example of the opposite trend, i.e., decreasing entropy, consider the use of *ing*-forms which diversify according to the analysis above shown for *filling* and *saturating*, i.e., they spread to different syntagmatic contexts. In the example in **Figure 10**, the terms in the cluster containing *assuming* exhibit a skewed frequency over time with decreasing entropy, reflecting in this case stylistic convergence, i.e., the tacit agreement on using particular linguistic forms rather than others. In particular, *assuming* has 30 close neighbors (including *supposing, assume, considering*) in the first decade, but only 13 close neighbors in the last decade, with *assuming, assume* dominating by frequency.

The effect of stylistic convergence on the reduction of the cluster entropy of *assuming* is visible also through a cursory look at some corpus concordances. Uses of *assuming* in the sense of "adopting" disappear (see example 1). Over time, *assuming* comes to be used increasingly at the beginning of sentences (example 2), the dominant use being the non-finite alternative to a conditional clause (*If we assume a/the/that...*). In terms of frequency, the dominant choice in the cluster is *assume*, presumably as a short form of *let us/let's assume* (example 3), a usage that is often associated with mathematical reasoning.

(1) *No notice is taken of any effervescence or discharge of air while it was assuming this color* (Cavendish, 1786).

(2) *Assuming a distribution of light of the form when x is the distance along the spectrum from the center of the line, the half breadth is defined as the distance in which the intensity is reduced to half the maximum* (Strutt, 1919).

(3) *Assume any three points a, b, c in the surface, no two of which are on one generator, [...]* (Gardiner, 1867).

5. SUMMARY AND FUTURE WORK

We have explored patterns of variation and change in language use in Scientific English from a diachronic perspective, focusing on the Late Modern period. Our starting assumption was that we will find both traces of diversification in terms of

TABLE 3 | Average cosine distance between function words vs. 2,000 randomly selected content words in the first and last decade of RSC 6.0 Open.

Group	Full vocabulary		Persistent vocabulary	
	1670	1920	1670	1920
Function words	0.44	0.51	0.46	0.47
Content words	0.45	0.70	0.44	0.63

To account for the constantly updated vocabulary of scientific terminology, we present both the results for all words in each decade (Full Vocabulary) and for only those words that appear in every decade (Persistent Vocabulary).

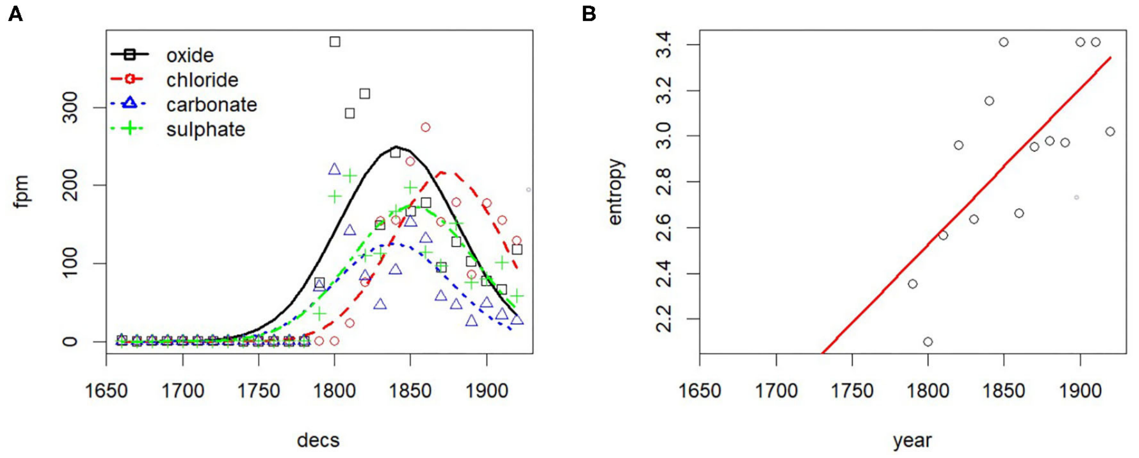


FIGURE 9 | Entropy increase on specific WE clusters signals terminological diversification. **(A)** Relative frequency. **(B)** Entropy.

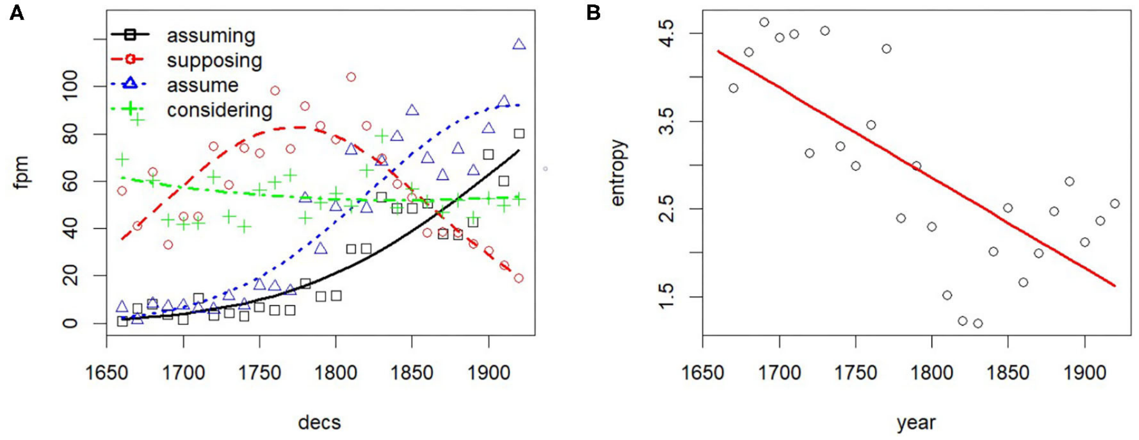


FIGURE 10 | Entropy decrease on specific WE clusters signals convergence in usage. **(A)** Relative frequency. **(B)** Entropy.

discourse field, thus pointing to register formation, as well as convergence in linguistic usage as indicator of an emerging scientific style. As a data set we used 250+ years of publications of the Royal Society of London [Royal Society Corpus (RSC), Version 6.0 Open].

We have elaborated a data-driven approach using three kinds of computational language models that reveal different aspects of diachronic change. Ngram models (both lemma and POS-based) point to an overall trend of consolidation in linguistic usage. But the lexical level dynamically oscillates between high peaks marking lexical innovation and lows marking stable linguistic use, where the peaks typically reflect new scientific discoveries or insights. At the grammatical level, we observe similar tendencies but at a much lower level and rate and the consolidation trend is much more obvious. Inspecting the specific grammatical patterns involved, we find that they mark what we commonly refer to as "scientific style," such as relational and passive clauses or specific nominal patterns for hosting terminology.

To investigate further the tendencies at the level of words, we have looked at aggregations of words from two perspectives—how words group together to form topics (development of fields of discourse as the core factor in register formation) and how specific words group together to form paradigms based on their use in similar contexts. Diversification is fully born out from both perspectives with glimpses of consolidation as well. Analysis on the basis of a diachronic topic model shows that topics diversify over time, indexed by increasing entropy over topic/word distributions, a clear signal of register formation. Analysis on the basis of diachronic word embeddings reveals that the overall paradigmatic organization of the vocabulary changes quite dramatically: the lexical space expands overall and it becomes more fragmented, the latter being a clear signal of diversification in word usage. Here, bursts of innovation are shown by increasing entropy on specific word clusters, such as terms for chemical compounds, mirroring the insights from lemma-based analysis with KLD. Also, patterns of convergence (confined uses of words) as well as obsolescence (word uses leaving the language) are shown by decreasing entropy on particular word clusters, such as the cluster of *ing*-forms. Taken together, we encounter converging evidence of diversification at different levels of analysis; and at the same time we find signs of linguistic convergence as an overarching trend—an emerging tacit agreement on "how to say things", a "scientific style."

In terms of methods, we have elaborated a data-driven methodology for diachronic corpus comparison using state-of-the-art computational language models. To analyze and interpret model outputs, we have applied selected information-theoretic measures to diachronic comparison. Relative entropy used as a data-driven periodization technique provides insights into overall diachronic trends. Entropy provides a general measure of diversity and is applied here to capture diversification as well as converging language use for lexis (word embeddings) overall and discourse fields (topic models) in particular.

In future work, we will exploit more fully the results from topic modeling and the word embeddings model of the RSC. For instance, we want to systematically inspect high and low-entropy word embedding clusters to find more features marking expansion (vs. obsolescence) and diversification (vs. convergence). Also, annotating the corpus with topic labels from our diachronic topic model will allow us to investigate discipline-specific language use (e.g., chemistry) and contrast it with "general" scientific language (represented by the whole RSC) as well as study the life cycles of registers/sublanguages. Especially interesting from a sociocultural point of view would be to trace the spread of linguistic change across disciplines and authors (e.g., Did people adopt specific linguistic usages from famous scientists?). Finally, we would like to contextualize our findings from an evolutionary perspective and possibly devise predictive models of change. Our results seem to be in accordance not only with our intuitive understanding of the evolution of science but also with evolutionary studies on vocabulary formation (e.g., Smith, 2004) showing how populations using specialized vocabularies are more likely to develop and take over when the selective ratio is pure efficacy in information exchange.

DATA AVAILABILITY STATEMENT

The Royal Society Corpus (RSC) 6.0 Open is available at: https://hdl.handle.net/21.11119/0000-0004-8E37-F (persistent handle). Word embedding models of the RSC with different parameter settings including visualization are available at: http://corpora.ids-mannheim.de/openlab/diaviz1/description.html.

AUTHOR CONTRIBUTIONS

YB curated the analyses on word embeddings showing the diachronic expansion of the scientific semantic space and lexical-semantic specialisation. SD-O carried out the analysis on diachronic trends in lexis and grammar using data-driven periodization and elaborating on features' contribution to change. PF trained the word embeddings and the topic models and designed and implemented the entropy-based diachronic analysis. ET provided the historical background and was involved in hypothesis building and interpretation of results.

FUNDING

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 232722074 - SFB 1102. The authors also acknowledge support by the German Federal Ministry of Education and Research (BMBF) under grant CLARIN-D, the German Common Language Resources and Technology Infrastructure. Also, the authors are indebted to Dr. Louisiane Ferlier, digitisation project manager at Royal Society Publishing, for providing advice and access to sources.

REFERENCES

Aitchison, J. (2017). "Psycholinguistic perspectives on language change," in *The Handbook of Historical Linguistics*, eds D. Joseph and R. D. Janda (London, UK: Blackwell), 736–743. doi: 10.1002/9781405166201.ch25

Aragamon, S. (2019). Register in computational language research. *Register Stud.* 1, 100–135. doi: 10.1075/rs.18015.arg

Atkinson, D. (1999). *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. New York, NY: Erlbaum. doi: 10.4324/9781410601704

Banks, D. (2008). *The Development of Scientific Writing: Linguistic Features and Historical Context*. London; Oakville, OM: Equinox.

Baron, A., and Rayson, P. (2008). "VARD 2: a tool for dealing with spelling variation in historical corpora," in *Proceedings of the Postgraduate Conference in Corpus Linguistics* (Birmingham).

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Biber, D., and Gray, B. (2011). "The historical shift of scientific academic prose in English towards less explicit styles of expression: writing without verbs," in *Researching Specialized Languages*, eds V. Bathia, P. Sánchez, and P. Pérez-Paredes (Amsterdam: John Benjamins), 11–24. doi: 10.1075/scl.47.04bib

Biber, D., and Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing. Studies in English Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511920776

Bizzoni, Y., Degaetano-Ortlieb, S., Menzel, K., Krielke, P., and Teich, E. (2019a). "Grammar and meaning: analysing the topology of diachronic word embeddings," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (Florence: Association for Computational Linguistics), 175–185. doi: 10.18653/v1/W19-4722

Bizzoni, Y., Mosbach, M., Klakow, D., and Degaetano-Ortlieb, S. (2019b). "Some steps towards the generation of diachronic WordNets," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19)* (Turku: ACL).

Blei, D. M., and Lafferty, J. D. (2006). "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, PA), 113–120. doi: 10.1145/1143844.1143859

Bochkarev, V., Solovyev, V. D., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* 11, 1–8. doi: 10.1098/rsif.2014.0841

Burke, P. (2004). *Languages and Communities in Early Modern Europe*. Cambridge: CUP. doi: 10.1017/CBO9780511617362

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195301571.001.0001

Cavendish, H. (1786). XIII. An account of experiments made by Mr. John McNab, at Henley House, Hudson's Bay, relating to freezing mixtures. *Phil. Trans. R. Soc.* 76, 241–272. doi: 10.1098/rstl.1786.0013

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). "No country for old members: user lifecycle and linguistic change in online communities," in *Proceedings of the 22nd International World Wide Web Conference (WWW)* (Rio de Janeiro). doi: 10.1145/2488388.2488416

Degaetano-Ortlieb, S., and Teich, E. (2016). "Information-based modeling of diachronic linguistic change: from typicality to productivity," in *Proceedings of the 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at ACL2016*, 165–173. doi: 10.18653/v1/W16-2121

Degaetano-Ortlieb, S., and Teich, E. (2018). "Using relative entropy for detection and analysis of periods of diachronic linguistic change," in *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018* (Santa Fe, NM), 22–33.

Degaetano-Ortlieb, S., and Teich, E. (2019). Toward an optimal code for communication: the case of scientific English. *Corpus Linguist. Linguist. Theory* 1–33. doi: 10.1515/cllt-2018-0088. [Epub ahead of print].

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). "Time-out: temporal referencing for robust modeling of lexical semantic change," in *Proceedings of the 57th Meeting of the Association for Computational Linguistics (ACL2019)* (Florence: ACL), 457–470. doi: 10.18653/v1/P19-1044

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). "Outta control: laws of semantic change and inherent biases in word representation models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen: Association for Computational Linguistics), 1136–1145. doi: 10.18653/v1/D17-1118

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE* 9:e113114. doi: 10.1371/journal.pone.0113114

Fankhauser, P., Knappen, J., and Teich, E. (2014). "Exploring and visualizing variation in language resources," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)* (Reykjavik), 4125–4128.

Fankhauser, P., Knappen, J., and Teich, E. (2016). "Topical diversification over time in the Royal Society Corpus," in *Proceedings of Digital Humanities (DH)* (Krakow).

Fankhauser, P., and Kupietz, M. (2017). "Visual correlation for detecting patterns in language change," in *Visualisierungsprozesse in den Humanities. Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu 2017)* (Zürich).

Fischer, S., Knappen, J., Menzel, K., and Teich, E. (2020). "The Royal Society Corpus 6.0. Providing 300+ years of scientific writing for humanistic study," in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)* (Marseille).

Fyfe, A., Squazzoni, F., Torny, D., and Dondio, P. (2019). Managing the growth of peer review at the Royal Society journals, 1865-1965. *Sci. Technol. Human Values*. 45, 405–429. doi: 10.1177/0162243919862868

Garcia, M., and Garćia-Salido, M. (2019). "A method to automatically identify diachronic variation in collocations," in *Proceedings of the 1st Workshop on Computational Approaches to Historical Language Change* (Florence: ACL), 71–80. doi: 10.18653/v1/W19-4709

Gardiner, M. (1867). Memoir on Undevelopable Uniquadric Homographics. [Abstract]. *Proc. R. Soc. Lond.* 16:389–398. Available online at: www.jstor.org/stable/112537

Gleick, J. (2010). "At the beginning: More things in heaven and earth," in *Seeing Further. The Story of Science and The Royal Society*, ed B. Bryson (London, UK: Harper Press), 17–36.

Goldberg, A. E. (2006). *Constructions at Work. The Nature of Generalizations in Language*. Oxford: OUP.

Görlach, M. (2001). *Eighteenth-Century English*. Heidelberg: Winter.

Görlach, M. (2004). *Text Types and the History of English*. Berlin, New York, NY: de Gruyter.

Grieve, J., Nini, A., and Guo, D. (2016). Analyzing lexical emergence in Modern American English online. *Engl. Lang. Linguist*. 20, 1–29. doi: 10.1017/S1360674316000526

Hall, D., Jurafsky, D., and Manning, C. D. (2008). "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI: Association for Computational Linguistics). doi: 10.3115/1613715.1613763

Halliday, M. (1985a). *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M. (1985b). *Written and Spoken Language*. Melbourne, VIC: Deakin University Press.

Halliday, M. (1988). "On the language of physical science," in *Registers of Written English: Situational Factors and Linguistic Features*, ed M. Ghadessy (London: Pinter), 162–177.

Halliday, M., and Martin, J. (1993). *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Cultural shift or linguistic drift? Comparing two computational models of semantic change," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Austin, TX). doi: 10.18653/v1/D16-1229

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *Int. J. Corpus Linguist*. 17, 380–409. doi: 10.1075/ijcl.17.3.04har

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hilpert, M., and Perek, F. (2015). Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguist. Vanguard* 1, 339–350. doi: 10.1515/lingvan-2015-0013

Hughes, J. M., Foti, N. J., Krakauer, D. C., and Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. U.S.A*. 109, 7682–7686. doi: 10.1073/pnas.1115407109

Hundt, M., Mollin, S., and Pfenninger, S. E. (eds). (2017). *The Changing English Language: Psycholinguistic Perspectives*. Cambridge, UK: CUP. doi: 10.1017/9781316091746

Hunston, S., and Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins. doi: 10.1075/scl.4

Jucker, A. H., and Taavitsainen, I. (2013). *English Historical Pragmatics*. Edinburgh: Edinburgh University Press.

Juola, P. (2003). The time course of language change. *Comput. Humanit*. 3, 77–96. doi: 10.1023/A:1021839220474

Jurish, B. (2018). "Diachronic collocations, genre, and DiaCollo," in *Diachronic Corpora, Genre, and Language Change*, ed R. J. Whitt (Benjamins: Amsterdam), 41–64. doi: 10.1075/scl.85.03jur

Kaiser, G. A., Butt, M., Kalouli, A.-L., Kehlbeck, R., Sevastjanova, R., and Kaiser, K. (2019). "Parhistvis: visualization of parallel multilingual historical data," in *Workshop on Computational Approaches to Historical Language Change* (Florence: Association for Computational Linguistics), 109–114.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., and Teich, E. (2016). "The Royal Society Corpus: from uncharted data to corpus," in *Proceedings of the 10th LREC* (Portoroz).

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). "Temporal analysis of language through neural language models," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (Baltimore, MD: Association for Computational Linguistics), 61–65. doi: 10.3115/v1/W14-2517

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141, 87–102. doi: 10.1016/j.cognition.2015.03.016

Klingenstein, S., Hitchcock, T., and DeDeo, S. (2014). The civilizing process in London's Old Bailey. *Proc. Natl. Acad. Sci. U.S.A*. 111, 9419–9424. doi: 10.1073/pnas.1405984111

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat*. 22, 79–86. doi: 10.1214/aoms/1177729694

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). "Diachronic word embeddings and semantic shifts: a survey," in *Proceedings of the 27th International Conference on Computational Linguistics (Coling)* (Sante Fe, NM: ACL), 1384–1397.

Leech, G., Hundt, M., Mair, C., and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511642210

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Ital. J. Linguist*. 20, 1–31.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). "Two/too simple adaptations of Word2Vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO: Association for Computational Linguistics), 1299–1304. doi: 10.3115/v1/N15-1142

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., and Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics* 41, 607–625. doi: 10.1016/j.poetic.2013.06.004

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 3111–3119.

Moskowich, I., and Crespo, B. (eds.). (2012). *Astronomy Playne and Simple: The Writing of Science between 1700 and 1900*. Amsterdam: Philadelphia, PA: John Benjamins. doi: 10.1075/z.173

Moxham, N., and Fyfe, A. (2018). The royal society and the prehistory of peer review, 1665-1965. *Historical J.* 61, 863–889. doi: 10.1017/S0018246X17000334

Nevalainen, T. (2006). "Historical sociolinguistics and language change," in *Handbook of the History of English*, eds A. van Kemenade and B. Los (London, UK: Wiley-Blackwell), 558–588. doi: 10.1002/9780470757048.ch22

Nevalainen, T., and Traugott, E. C. (eds.). (2012). *The Oxford Handbook of the History of English*. New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199922765.001.0001

Nguyen, D., Dogruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: a survey. *Comput. Linguist*. 42, 537–593. doi: 10.1162/COLI_a_00258

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Säily, T., Nurmi, A., Palander-Collin, M., and Auer, A. (eds.). (2017). "Exploring future paths for historical sociolinguistics," in *Advances in Historical Sociolinguistics* (Amsterdam: Benjamins). doi: 10.1075/ahs.7.01sai

Schmid, H. (1995). "Improvements in Part-of-Speech Tagging with an application to German," in *Proceedings of the ACL SIGDAT-Workshop* (Kyoto).

Schmidgen, H. (2011). *Das Labor/The Laboratory*. Europäische Geschichte Online/European History Online (EGO).

Smith, K. (2004). The evolution of vocabulary. *J. Theoret. Biol.* 228, 127–142. doi: 10.1016/j.jtbi.2003.12.016

Steyvers, M., and Griffiths, T. (2007). *Probabilistic Topic Models*. Hillsdale, NJ: Erlbaum.

Strutt, R. J. (1919). Bakerian lecture: A study of the line spectrum of sodium as excited by fluorescence. *Proc. R. Soc. Lond. A* 96:272–286. doi: 10.1098/rspa.1919.0054

Taavitsainen, I., and Hiltunen, T. (eds.). (2019). *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*. Amsterdam: Benjamins. doi: 10.1075/z.221

Taavitsainen, I., and Jucker, A. H. (eds.). (2003). *Diachronic Perspectives on Address Term Systems*. Amsterdam: Benjamins. doi: 10.1075/pbns.107

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to diachronic conceptual change. *arXiv[Preprint]a.rXiv:1811.06278*.

Thoiron, P. (1986). Diversity index and entropy as measures of lexical richness. *Comput. Humanit*. 20, 197–202. doi: 10.1007/BF02404461

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). "Rethinking LDA: why priors matter," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, BC: Curran Associates, Inc.), 1973–1981.

Xu, Y., and Kemp, C. (2015). "A computational evaluation of two laws of semantic change," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci)* (Pasadena, CA).

Yang, T.-I., Torget, A., and Mihalcea, R. (2011). "Topic modeling on historical newspapers," in *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (Portland, OR).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/frai.2020.536086

Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data

*Christoph Purschke**

Department of Humanities, Institute for Luxembourgish Language and Literature, University of Luxembourg, Esch-sur-Alzette, Luxembourg

OPEN ACCESS

Edited by:
David Jurgens,
University of Michigan, United States

Reviewed by:
Alberto Barrón-Cedeño,
University of Bologna, Italy
John Bellamy,
Manchester Metropolitan University,
United Kingdom

***Correspondence:**
Christoph Purschke
christoph.purschke@uni.lu

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *18 February 2020*
Accepted: *02 September 2020*
Published: *22 October 2020*

Citation:
Purschke C (2020) Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data.
Front. Artif. Intell. 3:536086.
doi: 10.3389/frai.2020.536086

Attitudes are a fundamental characteristic of human activity. Their main function is the situational assessment of phenomena in practice to maintain action ability and to provide orientation in social interaction. In sociolinguistics, research into attitudes toward varieties and their speakers is a central component of the analysis of linguistic and cultural dynamics. In recent years, computational linguistics has also shown an increased interest in the social conditionality of language. To date, such approaches have lacked a linguistically based theory of attitudes, which, for example, enables an exact terminological differentiation between publicly taken *stances* and the assumed underlying *attitudes*. Against this backdrop, the present study contributes to the connection of sociolinguistic and computational linguistic approaches to the analysis of language attitudes. We model a free text corpus of user comments from the RTL.lu news platform using representation learning (*Word2Vec*). In the aggregated data, we look for contextual similarities between vector representations of words that provide evidence of stances toward multilingualism in Luxembourg. We then contrast this data with the results of a quantitative attitudes study, which was carried out as part of the crowdsourcing project "Schnëssen." The combination of the different datasets enables the reconstruction of socially pertinent attitudes represented in public discourse. The results demonstrate the central importance of attitudes toward the different languages in Luxembourg for the cultural self-understanding of the population. We also introduce a tool for the automatic orthographic correction of Luxembourgish texts (*spellux*). In view of the ongoing standardization of Luxembourgish and a lack of rule knowledge in the population, orthographic variation—among other factors like code-switching or regional dialects—poses a great challenge for the automatic processing of text data. The correction tool enables the orthographic normalization of Luxembourgish texts and with that a consolidation of the vocabulary for the training of word embedding models.

Keywords: computational sociolinguistics, attitudes, crowdsourcing, low-resource languages, Luxembourgish, multilingualism, orthographic normalization, representation learning

INTRODUCTION

Attitudes toward language and other cultural phenomena are one of the basic characteristics of social practice. They play a central role for the way people use, perceive, and evaluate language. For example, the assessment of a social style or regional variety (e.g., as opposed to the standard variety) in a specific situation has an impact on behavior in competitive situations (Heblich et al., 2015). The same holds true for how people perceive other people in terms of character traits or other aspects of social interaction (Kristiansen, 2009). Attitudes arise in practice in the form of "relevance-driven targeting and evaluation routines [...] that sediment in an individual's stock of knowledge and are situationally (re)constructed in interaction" (Purschke, 2015, p. 49). Attitudes are therefore routinized *judgments* about phenomena in everyday life, which can become apparent in interaction in the form of *stances* (Jaffe, 2013), that is, in speech acts or other types of action. However, there is no demonstrable direct link between a person's attitudes and their actions (Soukup, 2012). The reason for this lies in the diverse implicit and explicit, self-related and social *norms* that determine social interaction, and, therefore, the emergence, structuring, and externalization of attitudes. For example, not every attitude is *socially appropriate* in every situation, such as politically controversial opinions when talking to a superior at work. In addition, not all attitudes are equally *cognitively accessible* and *consciously controllable* with regard to their appearance in and relevance for action (Pharao and Kristiansen, 2019). As a consequence, we have to take into consideration different aspects of the *cognitive organization*, *social embedding*, and *practical functions* of attitudes, for example, the complex relation between the long-term stability of many attitudes (e.g., prejudice against certain dialects; Preston, 2015), their general changeability through new experiences (e.g., through direct contact with speakers of a stigmatized variety; Giles and Marlow, 2011), and their situational expression in concrete interactions (e.g., the use of dialectal features as stance markers in chat communication; Tophinke and Ziegler, 2014).

Research on attitudes dates back to the early days of psychology and has been a topic of long-standing tradition in the humanities and social sciences. In sociolinguistics, attitudes have been examined with a wide range of methodological approaches and against the backdrop of different theoretical frameworks. Albarracín and Johnson (2018) provide a good overview about the development of the field. Regardless of methodological and theoretical discussions about how to describe and survey attitudes best, it has been shown that and to what extent attitudes are important for the practical organization of social interaction. For example, in the German speaking area, the perception and evaluation of linguistic variation is closely related to the overall dynamics of regional dialects, and this connection derives directly from the sociocultural orientation of the language users (Purschke, 2018). In addition, people's attitudes toward language in general and the different varieties present in a speech community substantially influence their migration behavior (Lameli et al., 2015).

At the same time, this close connection between language use and language evaluation poses one of the biggest challenges to the computational processing and modeling of language in computational linguistics (Hovy, 2018). Basic traits of language practice, such as social meaning, irony, mimicking citation, and other forms of stylization cannot reliably be detected and processed by algorithms (e.g., in tasks like sentiment analysis, machine translation, or chat bots). Furthermore, models and algorithms work best for standardized datasets in high-resource languages and seem to reproduce aspects of demographic and social bias in automated processing (Garimella et al., 2019). As a consequence, the applicability and appropriateness of many NLP applications for everyday language is still limited, despite the great advances in computer science and AI research (Bender and Koller, 2020). In recent years, there has been a new trend in the NLP community that is increasingly concerned with language as a social phenomenon and that tries to incorporate sociolinguistic knowledge into the analysis of data and the development of new tools and models (Broadwell et al., 2013; Eisenstein, 2015; Nguyen, 2017; Purschke and Hovy, 2019).

This article is committed to the same goal. The aim of the text is to reconstruct language attitudes toward multilingualism in Luxembourg with the help of different data types. On the one hand, we aggregate stances toward language and multilingualism in free text data and evaluate them using computational linguistic methods. We then compare the data with the results of a sociolinguistic questionnaire survey that was carried out with the help of a mobile crowdsourcing application. A comparison of the different data types shows that attitudes can be successfully reconstructed from free text data and that the patterns found reflect the attitudes of people toward multilingualism in Luxembourg as well as certain aspects of public discourse. In terms of methodology, the text thus makes a contribution to the field of computational sociolinguistics by trying to systematically relate computational linguistic and sociolinguistic approaches in analysis. From a theoretical point of view, the article provides proof of the importance of *contextual knowledge* for the organization of social practice, with a special regard to the role of attitudes as *situated evaluation routines*. Furthermore, the article contributes to the development of computational linguistic resources for Luxembourgish as a low-resource language, that is, the automatic normalization of orthographic and regional variation in text data.

MULTILINGUAL LUXEMBOURG

The sociolinguistic setting in Luxembourg is comparably complex. It has developed as a result of a fickle history in contact with neighboring cultures (especially France and Germany). In addition, socio-economic migration, the country's specialization in the private financial industry, and the presence of several European institutions play an important role in the emergence and dynamism of the current language regime. With a total population of 613,000, the Grand Duchy has a very high proportion of foreign residents of 47.5%. In addition, there are 192,000 cross-border commuters coming in from Germany, France, and Belgium every day (STATEC, 2019). The country has three official languages, the national language Luxembourgish, and French and German as administrative languages. Luxembourgish multilingualism is also characterized by strong minority languages (Portuguese, Italian) and an increasing presence of English. Language use and the social embedding of the different languages in Luxembourg are organized on a domain-specific basis (Erhart and Fehlen, 2011). For example, French serves as the language of legislation and jurisdiction, but debates in Parliament take place in Luxembourgish. The print media are traditionally dominated by German (and to a limited extent French), while radio and national television broadcast largely in Luxembourgish. German is the language of alphabetization, but the school system as a whole is also designed to promote multilingualism, with a strong copresence of French. Luxembourgish is the language of everyday communication among Luxembourgers and has undergone processes societal and political *Ausbau* in the past 15 years (Gilles, 2019), which have resulted, among other things, in a new law promoting Luxembourgish in 2018, by means of which its societal anchoring is to be strengthened. The language has developed into a written variety that is suitable for all communicative occasions, from informal communication in social media to public inscriptions and formal letters, and the official orthography has been consolidated and modernized in 2019. At the same time, the majority of the population does not have an in-depth knowledge of the official spelling rules, because Luxembourgish is not an integral part of the school curriculum.

Given its sociocultural diversity and strong demographic dynamics (the population has grown by 39.7% since 2001; STATEC, 2019), the language regime is currently on the move. While Luxembourgish is increasingly present in all social domains, the role of German as a bridge language (traditionally seen as "written Luxembourgish"; Kloss, 1952) is clearly decreasing. At the same time, the importance of French is increasing, above all because of the high proportion of foreign employees in the private sector. Additionally, the social presence of English is increasing due to its growing importance for tourism, economy, and pop culture. While French traditionally plays the role of a cultural prestige language, the young generation in particular shows a clear preference for English (and indirectly German due to its close relationship with Luxembourgish). Multilingualism and especially the societal role of Luxembourgish have been frequent topics in public debates in recent years (Horner and Weber, 2008; Garcia, 2014). Following a referendum on the right to vote for foreigners in 2015 and an increasing politicization of language in public discourse and political action, the discussion about the languages of the country has developed into a "replacement discourse." In this context, languages serve as a proxy for societal disputes, for example, the demographic development, rising living costs, and democratic legitimation of politics. Many of these topics can also be found in discussions on social media (especially Facebook) and in the user comments of the RTL.lu news platform (*Example 1*).

> **Example 1: Language-related comment from the RTL data set**
> *Et soll endlech klip % klor gesetzlich verankert gin das jus nach L hei emgangssproch ass, d.h wen well hei schaffen op brout verkafen oder deck plaz op da bank MUSS L kennen. Dat muss dach meiglich sin* [2016-02-21].
> **Translation:** It should finally be anchored in the law that Luxembourgish is the only colloquial language here, which means that anyone who wants to work here, whether selling bread or a fat job in the bank, must be able to speak Luxembourgish. That must be possible.

In this example, the author takes a clear stance on the language regime by demanding Luxembourgish as the only colloquial language for the country. They combine this with a demand for linguistic integration from foreign workers. In addition to the close connection between linguistic and societal issues in public discourse, the comment also illustrates some of the challenges in dealing with Luxembourgish text data: The text contains many spelling mistakes (e.g., *jus* instead of *just* "just, only," *emgangsproch* instead of *Ëmgangssprooch* "colloquial language"), irregular use of capitalization and punctuation, abbreviations like *L* for *Lëtzebuergesch* "Luxembourgish," and colloquial expressions. This variability poses a particular challenge for automated text processing, especially because of the large amount of orthographic variation.

Against the backdrop of the complex and dynamic Luxembourg multilingualism, the aim of the present study is to examine the attitudes of the population toward multilingualism and the role of Luxembourgish in particular. On the one hand, the analysis is based on user comments from the RTL.lu news platform, on the other hand, answers from a sociolinguistic questionnaire survey on attitudes toward multilingualism are taken into consideration.

DATA AND METHODS

In the following section, the different data sources are discussed. This involves the respective characteristics of the data, but also their preparation and modeling for the subsequent analysis. First, we present the user comments from RTL.lu. In this context, we discuss the particular challenges when working with Luxembourgish text data that require a special preprocessing workflow. In a second step, we discuss the questionnaire data. Since these data stem from a crowdsourcing project, certain preprocessing steps are also necessary in this case.

Mining Attitudes From RTL.lu User Comments

Dataset

The data for the computational linguistic analysis stem from the RTL.lu news platform. The RTL media group is the largest news provider in the country and has television and radio programs as well as a widely used online news portal. The platform has existed since 2008 and is the only news offering to date that is entirely in Luxembourgish. As part of a project to develop semantic annotation algorithms for Luxembourgish text data at the University of Luxembourg ("STRIPS" project; Gierschek et al., 2019), RTL has made all the articles published on the platform as well as the associated user comments available for research. The project primarily uses the data to measure sentiment in user comments. In addition, the data can also be used for the investigation of orthographic variation (temporal development of correctness and individual norm accommodation) or for discourse analytical questions, for example, the reconstruction of language attitudes.

The dataset comprises a total of 179,298 news articles and 585,358 user comments from the period between 2008 and 2018. All comments are anonymous and, in addition to a time stamp, contain information about the article to which they refer. Thematically, the corpus covers the entire range of topics offered on the media platform: national and international news, topics from society, culture, and science, sports, local journalism, but also reader contests or reports. The majority of the texts are written in Luxembourgish. While the news articles are largely spelled correctly orthographically, the user comments show diverse sources of linguistic variation:

- *correctness:* Since the development of Luxembourgish as a written variety has taken place over the past 15 to 20 years and its standardization has not yet been completed, the early contributions tend to show a greater orthographic variation than more recent contributions, especially with regard to their correctness. In view of the lack of social anchoring of the official rules in the population, however, the recent contributions are also very variable orthographically.
- *formality:* The comments express a range of textual formality, from some early comments similar to letters (with a salutation and signature) to informal texts typical for online communication that are conceptually largely based on oral language.
- *mediality:* The texts show the expected range of medium-specific writing resources that are typical for digital writing. This includes variable use of upper- and lower-case letters, the use of emoji and acronyms, irregular punctuation, or onomatopoetic writing to express emphasis.
- *regionality:* In addition to orthographic variation, the texts are also characterized by regional variation. Although extensive processes of dialect leveling have already taken place in Luxembourgish, there are still diverse traces of regional spellings in the texts, e.g., forms like *wuar* or *woar* for *war* "was".
- *multilingualism:* While the majority of the contributions is written in Luxembourgish, the multilingual competence of the writers results in many texts that contain elements of code-switching into German, French, or English. In addition, there is the characteristic of Luxembourgish as a "hybrid" language, that is structurally close to German and at the same time has integrated many elements from French.

These characteristics of online writing are not exclusive to Luxembourgish. In fact, we find some of them (correctness, regionality) in many smaller languages that have not been (fully) standardized, while others (formality, mediality) are typical for (the development of) online writing in general, as is code-switching in multilingual communities. However, the combination of the different characteristics, combined with the comparatively good availability of machine-readable data, represents a special feature of Luxemburgish as a research topic. Additionally, the Luxembourgish writing system has some systemic peculiarities, for example, there is a contextual (phonetic) rule according to which the endings *-n* or *-nn* are not to be written before initial vowels and some consonants in the following word, the so-called "n-rule" (Zenter für d'Lëtzebuerger Sprooch, 2019).

In the following, we analyze the RTL user comments as for language attitudes. We use the articles only as a supplementary data source for preprocessing (i.e., learning of an additional embedding model for orthographic normalization). In a follow-up study, it would be worthwhile to look for systematic connections between journalistic reporting and user discussions.

Preprocessing

In view of the extent of linguistic variation, we develop a special preprocessing workflow for the user comments. The goal is to reduce the amount of variant spellings for lemmas in the data in order to obtain a smaller and semantically consolidated vocabulary for the analysis. The workflow includes cleaning the texts from special characters and markup language, sorting out non-Luxembourgish contributions through language detection, tokenizing the data, and orthographic normalization. We implement all work steps in *Jupyter Notebooks* with *Python 3*.

Cleaning of the data

Due to the origin of the texts (online news portal) and the period of their creation (2008–2018), the texts first have to be cleaned of special characters, incorrect encodings, and markup language. In addition, since its foundation, the news platform has undergone several changes in the technical basis, which are reflected in the data in the form of different markup standards. As a consequence, data cleaning has to deal with the removal of html tags and other markup elements for online texts, the conversion of various text encoding standards into Unicode characters, and also the removal of special characters and hyper-text content (links and other embedded elements). In order to find a tailored solution for the many encoding errors in the data, we use a dictionary-based approach to replace these characters.

Language detection

In a second step, we process all comments with the help of the package *langdetect* to identify the text language. For this purpose, we train a language profile for the recognition of Luxembourgish on the basis of the RTL news articles and implement it into the package. In this way, we can separate the Luxembourgish texts from comments in other languages. However, the recognition only works reliably on the comment level[1]. This preprocessing

[1] Detection accuracy was tested manually using a random sample of 1,000 texts labeled as Luxembourgish (100% correctly identified). Identification of non-Luxembourgish texts gives mixed results: Overall, accuracy is 64% for a random sample of 1,000 texts. Texts with wrong labels mainly concern very short texts

step reduces the amount of comments for the analysis to 544,143 posts. It also reduces the influence of multilingualism in the data. However, better language models are needed to process phenomena such as code-switching and loan vocabulary on the sentence level. For the further steps, this means that a certain number of foreign language elements remain in the text corpus (most of these words are filtered out by the frequency threshold during the training of word the word embedding model, though).

Tokenization

We then tokenize the data using the package *spaCy*. Since November 2019, this package has language support for Luxembourgish, including tokenization and POS tagging[2]. Compared to other resources (Sirajzade and Schommer, 2019), processing in *spaCy* works reliably for tasks like POS tagging, makes use of state-of-the-art algorithms and data formats, and also takes peculiarities of Luxembourgish spelling into consideration, such as the correct separation of *d'* as a definite article in words like *d'Saach* "the thing, the matter."

Orthographic normalization

The most challenging step in data preparation is the orthographic normalization of the data. In view of the diverse sources of linguistic variation, we introduce the *Python* package *spellux*[3], a pipeline that helps reducing the number of spelling variants in the corpus without having to exclude them for the subsequent training of a word embedding model (i.e., by setting a frequency threshold parameter). For this purpose, we use a multi-stage process, which compares a variant with different correction instances and, in unambiguous cases, corrects the text. Different correction resources are available for this task:

- *Word embedding model*: Based on the entire corpus, that is, user comments and news articles, we train a vector space model using the *gensim* package (word embedding with *Word2Vec*; Mikolov et al., 2013). The goal is to use representation learning to identify orthographically similar forms of the same lemma with the model. This is possible because word embedding models structure corpora in a high-dimensional vector space according to the *contextual similarity* of words based on semantic-syntactic co-currencies. The use of all data for the embedding model makes it possible to compare the individual spelling variants in the comments with the correct spellings in the articles—because they appear in comparable contexts in terms of linguistic structure. We use the following common hyperparameters to train the model (Mikolov et al., 2013; Pierrejean and Tanguy, 2018): dimensions: 200, window size: 5, iterations: 5, word frequency threshold: 25, downsampling of frequent words: 1e−3.
- *Correction dictionary*: We implement a list of lemmas and spelling variants from the online correction tool "spellchecker.lu." With the help of this tool, writers can check Luxembourgish texts online and replace spelling mistakes with correct variants. The entered variants and correction lemmas are logged in the tool. We create a correction dictionary from these, which contains the most frequent (f > 20) spelling variants for each lemma as well as the summary correction frequency for all variants of a lemma (Note that this dictionary is only used for training the correction models in *spellux* and not part of the official release).
- *tf-idf matrix*: We train a tf-idf correction matrix using the entire lemma list from the correction dictionary as a basis, and the *TfidfVectorizer* method in the package *scikit-learn*. In doing so, we determine the k-nearest neighbor for a given variant in the lemma list.
- *Norvig spelling corrector*: Additionally, we implement an adaptation of the spelling corrector by Peter Norvig that has been tailored to Luxembourgish orthography[4]. The corrector evaluates the most likely correction candidate for a given variant based on a large text sample (of RTL news articles).

For orthographic normalization, we use the following workflow:

- First, we compare each word form with the lemma list in the correction dictionary. We classify variants recorded as lemma as correct (including some false positives for homographic forms).
- Second, we check whether forms that are not included in the lemma list are listed as spelling variants in the correction dictionary. If the form is recorded as a variant of exactly one lemma, we replace it with the corresponding lemma in the text. In cases where a form is used as a spelling variant for several lemmas (e.g., *as* for *ass* "is" and *als* "as"), we run an extended correction routine. To do so, we can choose from different correction modes (see the package documentation for further details): We can either check a variant for its vector similarity ($cos\ \theta$) with all words in the word embedding model to determine a correction candidate by its contextual similarity with the variant, we can determine a candidate using the tf-idf matrix, we can evaluate a candidate using the Norvig corrector, or can we use a combination mode that accesses all three correction modes. To assure correction accuracy, and for best candidate evaluation in the combination mode, we evaluate the string similarity of correction candidates against the input form using the *Jaro Winkler* distance measure in the package *jellyfish.* In the event of a good enough match, we replace the variant with the best candidate. Given that the word embedding model was trained on the entire RTL corpus, we choose the embedding model as the default correction mode.
- If we cannot determine a clear candidate using the correction routine, the spelling variant is not corrected.
- We write each pair of spelling variant and lemma found to a dynamic matching dictionary to save the matches for later occurrences of the same variant and speed up text correction.

The comment corpus comprises 38,568,920 words. Through the orthographic normalization and case conversion, we reduce

that do not contain much language-specific content, or texts with a lot of code-switching. If we only consider texts with a length of more than 200 characters, the recognition rate increases to 96% for non-Luxembourgish texts.

[2] Language support for Luxembourgish in *spaCy* has been developed by the author and Peter Gilles.

[3] https://github.com/questoph/spellux/

[4] https://norvig.com/spell-correct.html

the number of unique words in the corpus from 1,102,377 to 1,017,175. Nevertheless, there are 680,300 unique words in the corpus for which we find no replacement using the available correction resources. Some of these are misspellings that are not yet recorded in the correction dictionary, some are words that are missing from the lemma list, some stem from foreign language material left in the comments (code-switching, citations). Further processing would be necessary for these words to improve the automatic normalization of the texts, for example, the semi-automatic extension of the correction dictionary by these variants.

Modeling

On the basis of the orthographically normalized texts, we train a new word embedding model (using the same training hyperparameters as before) that includes only the user comments. This model serves as the basis for the reconstruction of language attitudes toward multilingualism. According to the logic behind representation learning, the vectors of words that have a closer semantic-syntactic connection should have a higher contextual similarity in the vector space model. For example, in the data, the country name *Lëtzebuerg* "Luxembourg" is contextually more similar to the vector representation of its polity (*Monarchie* "monarchy," cos $\theta = 0.260223$) than to the vector for the word "democracy" (*Demokratie*, cos $\theta = 0.245135$)—nevertheless, Luxembourg is of course a democratically governed country. However, we cannot interpret this relation as an exact representation of the semantic-syntactical closeness of the concepts in question. For example, the vector for *Diktatur* ("dictatorshop," cos $\theta = 0.273865$) is even closer to *Lëtzebuerg*.

Nevertheless, it is possible to interpret the contextual similarity of word vectors in the embedding model as statements about the relative *discursive proximity* of concepts in the dataset, for example, regarding language attitudes. Words whose vector representations are closer together in the model are more likely to appear in similar semantic and syntactic contexts—without necessarily specifying the exact quality of this relation. That is why we are interpreting this relation holistically, that is, as a combination of semantic (*concept similarity*) and syntactic (*context similarity*) information that, in sum, mirrors the sociopragmatic use of a word relative to others in the corpus. To avoid false conclusions, however, and given the general vulnerability of word embedding models to input variability and training hyperparameters, we compare the data with the results of a questionnaire survey on language attitudes. The comparison of the learned word representations and the empirically tested language attitudes makes it possible to draw conclusions about the representation and evaluation of languages in discourse, but also about the meaningfulness of the learned representations for the analysis.

Related Research

The general benefit of representation learning and distributional semantics for the reconstruction of the social meaning of concepts has already been examined in computational linguistics. Grondelaers and Speelmann (2015) use vector space models to cluster keywords returned in a free-response experiment on language attitudes into semantically meaningful dimensions for interpretation. Garg et al. (2018) demonstrate how the temporal encoding of word embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities. And Kozlowski et al. (2019) show that vector representations of semantic word relations in such models (e.g., for *man—woman*, *rich—poor*) can be related to common cultural stereotypes in public discourse. In addition, there are other approaches for determining attitudes and emotions in language data.

For example, Dong et al. (2019) show based on crowdsourced questionnaire data that the cross-cultural perception of social roles differs considerably and that these differences can be predicted using attributive descriptors or associated actions for social roles in context. Hassan et al. (2010) introduce a method to identify reciprocal attitudes of participants in an online discussion forum by evaluating positive or negative elements in sentences. The approach is expanded to the "AttitudeMiner" system in Abu-Jbara et al. (2012). Dasigi et al. (2012) automatically detect subgroups of users in online discussion threads based on implicit attitudes expressed by similar language use, similar to Somasundaran and Wiebe (2009) who focus on debate genre and opinion-based social stance in multiauthor threads. Rodríguez-Penagos et al. (2012) introduce a modular and scalable framework for opinion mining in social media data based on posts about Spanish telephone services and products. Lin et al. (2013) automatically track discussion dynamics in social media using topic-based attitude modeling and topical position mapping to determine the participants positionings toward each other. And Chuang and Hsieh (2015) perform a binary classification task to determine stances in social media posts with a lexicon-based approach that makes use of linguistic feature analysis and manual annotation.

There are also a number of earlier studies that employ different methods to try to determine the contextual emotional value of sentences in text data, be it with the help of keyword matching techniques (Chuang and Wu, 2004; Strapparava et al., 2007), calculations of emotion points (Taboada and Grieve, 2004), sets of linguistic interpretation rules (Boucouvalas, 2003; Chaumartin, 2007), sets of predefined attitude labels (Neviarouskaya et al., 2009), or machine learning methods (Aman and Szpakowicz, 2008; Strapparava and Mihalcea, 2008). Pang and Lee (2008) offer a comprehensive overview of early work on sentiment analysis and opinion mining.

So far there is hardly any comparable work for Luxembourgish, as well as for attitudes toward multilingualism in general. As part of the STRIPS project (Gierschek et al., 2019), we are currently developing an engine for automatic sentiment analysis for Luxembourgish. The system makes use of manually annotated training data, word embedding, and recursive neural networks for sentiment prediction.

What is striking about most computational linguistic work on the nexus *ideology—attitude—stance –sentiment—emotion* is the lack of a coherent conceptual basis that is grounded in linguistic and socio-psychological theory, and with it a clear delimitation of the different concepts involved (see for example Munezero et al., 2014). Often the terms for the examined concepts change several times within the same text. In this respect, the present study may also serve as a contribution to the theoretical foundations of computational sociolinguistics with regard to the social meaning of linguistic phenomena in interaction. In many studies, there is also a problematic equation of observable language use (i.e., stance, sentiment) and the assumed underlying cognitive entities (i.e., attitude, emotion), while the social-psychological literature on attitudes particularly emphasizes the lack of a direct attitude-action link. In addition, many studies seem to be primarily interested in the technical aspects of the implementation, prediction accuracy, and evaluation of methods for the detection of emotions or opinions in utterances, less in their applicability to and meaningfulness for sociolinguistic research. Against this backdrop, the combination of different data types for the purpose of a sociolinguistic analysis of attitudes is particularly worthwhile.

Crowdsourcing Attitudes With the "Schnëssen" App

Dataset

The data for the sociolinguistic analysis stem from a questionnaire survey as part of the crowdsourcing project "Schnëssen" (Entringer et al., forthcoming). The project is an initiative of the Institute for Luxembourgish Language and Literature at the University of Luxembourg and aims to document variation and change in present-day spoken Luxembourgish. For this purpose, we have developed a mobile research app with which speakers of Luxembourgish can record their own language use. Since 2018, we have collected voice data from more than 2,500 speakers and for more than 500 linguistic phenomena in this way. In addition to the language survey, a sociolinguistic questionnaire can also be accessed via the app, which specifically asks about the participants' attitudes to multilingualism and Luxembourgish. We use a specially developed quantitative instrument to collect the attitudes.

Participants are asked to rate comments on five-tier Likert scales. In contrast to comparable studies, we take care to ensure that the statements to be assessed mirror situations that respondents are familiar with and encounter frequently in everyday life. A general weakness of quantitative attitude measurements should be avoided in this way (see Purschke, 2014 for a discussion): Comparable studies often ask about abstract concepts or assessments for which there is no direct correspondence in the everyday experience of the respondents. As a consequence, in many cases, the respondents must first form an opinion to the subject of the question instead of activating their existing everyday knowledge.

The questionnaire covers four thematic areas: the development of multilingualism in the country, the state of Luxembourgish, the social presence of the most important languages, and individual language preferences in everyday situations. Between April and January, 2019, 2,158 complete questionnaires have been collected that can be used for the analysis. In addition, each participant has created a social profile in the app that contains the most important biographic and linguistic information. This includes language skills, places of residence, stays abroad, educational profile, age, and gender. In view of the technical and linguistic requirements of the app, the data shows a characteristic demographic bias: The app is entirely in Luxembourgish and also requires knowledge of German and French for translation tasks. As a consequence, the app has linguistic preconditions that are primarily met by Luxembourgish native speakers, who make up more than 90% of the sample, whereas the other half of the population is hardly represented. In addition, there is the usual demographic bias for app-based surveys that rely on voluntary work, that is, young, well-educated, female participants are overrepresented in the sample (Behrend et al., 2011).

Preprocessing

In order to prepare the data for analysis, we have to match the questionnaire data with the users' social profiles (using a device-specific unique identifier). The reason for this lies in the fact that the questionnaire is embedded in the app as an independent task, but the creation of a social profile is only mandatory for the app's recording function. As a consequence, many participants filled out the questionnaire without creating a social profile. In addition, there are cases in which several people made recordings or filled out the questionnaire using the same device, which is why sometimes there are several social profiles and only one questionnaire for the same universal identifier and vice versa. To deal with this situation, we first match the unique questionnaires and unique social profiles. The remaining cases of doubt, in which the number of social profiles and questionnaires differ, we match manually if possible. After preprocessing, 1,832 completed questionnaires remain, which can be assigned to a unique social profile. These data form the basis for the following analysis.

Related Research

So far, there are only a few studies on attitudes and stances toward Luxembourg multilingualism. These focus primarily on the language preferences of speakers in various everyday situations, for example, in work contexts or leisure activities (Fehlen, 2009; Fehlen and Heinz, 2016). The studies show a clear connection between (first) language competence and language preference in practice. In addition, the practical requirements of everyday life play a central role in the situational choice of a language. Conrad (2017) includes similar questions in the analysis of contact-related variation in Luxembourgish to explain the preference of the speakers for Germanic or Romance variants in use. Redinger (2010) deals with language attitudes and language behavior in the Luxembourg educational system in combing a questionnaire survey with an ethnographic investigation of in-class code-switching. Wagner (2012, 2013) investigates writing strategies and their relation to language use and ideologies in social media discussions on Facebook. In a similar vein, Belling and de Bres (2014) investigate the role of Luxembourgish for group negotiations and identity constructions in a multilingual Facebook group. Language ideologies and the practical negotiation of multilingualism in the workplace, with particular attention to cross-border commuters, are the focus of the studies by Franziskus (2013), De Bres (2014), and De Bres and Franziskus (2019). Lately, Bellamy and Horner (2018) focus on ideological positionings in interaction with regard to the societal role and linguistic status of

Luxembourgish as a national language. In a questionnaire survey with more than 2,000 participants, Stölben (2019) examines the Luxembourgers' attitudes toward the official languages in the country, with a special focus on German. The study documents the complex attitudinal horizon of the Luxembourgers regarding the different languages in the country, with both the domain-specific organization of multilingualism as well as individual factors such as language competence and social environment contributing to individual attitudes.

All studies establish a clear connection between language competence, language preference, and sociocultural orientation in everyday life. The role of Luxembourgish as a practical means of individual social positioning (*identity level*) and a symbolic resource of group-related identification (*ideology level*) is particularly important in this context. For the study of language attitudes, this means that the position of Luxembourgish in the complex Luxembourg multilingualism is crucial, but also the structure and dynamics of the language regime as a whole.

RESULTS

Based on these findings, we present selected results of the questionnaire survey below and contrast them with queries to the word embedding model trained on the user comments. Since the comments are free text data that represent reactions to journalistic content, many texts contain clear positive and negative stances on certain topics that seem suitable for the aggregating reconstruction of attitudes. *Example 2* gives another example of such public stances in the dataset that also illustrates the difference between *explicit* and *implicit* aspects of stances and attitudes in practice: first, the author explicitly positions themselves in favor of Luxembourgish by calling for resistance (*Fannen och mir mussen ons wiehren.* "I also think we must fight back."), followed by a direct call for action (*Rett ons sproch* "Save our language". Then, in addressing the audience they code-switch from Luxembourgish to English (*be united people*). In view of the language-ideological subject of the comment, the switch to English is likely to take place at an implicit level of stance-taking, also because code-switching is part of the highly routinized repertoire of multilingual speakers in Luxembourg.

> **Example 2: Language-related comment from the RTL data set**
> *Fannen och mir mussen ons wiehren. Rett ons sproch, be united people* [2016-02-21]
> **Translation:** I also think we must fight back. Save our language, be united people

The results of the computational text analysis are not to be equated with the quantitatively surveyed attitudes in the questionnaire, though. By comparing the two datasets, however, we can draw conclusions concerning attitudes toward multilingualism present in the Luxembourg population. Comments and survey data serve as complementary data sources that link publicly taken stances in discourse to underlying attitudes that impact the structure and dynamics of the language regime in the country. For example, the growing discussion about the societal role of Luxembourgish in recent years has had a direct impact on politics, which was reflected in the issue of language as a topic in the national election campaign in 2018 as well as in the newly introduced language promotion law for Luxembourgish. Connecting these two datasets is the particular challenge—and the particular contribution—of the following computational sociolinguistic analysis.

The Social Presence of Languages in the Language Regime

The first set of results relates to the social presence of the various languages in the country, that is, their position and symbolic value in the language regime. There are a couple of questions in the questionnaire that are of interest in this context. This includes the question of which of the most important languages "belong" to the country (**Table 1**). So, the question is about the cultural self-image of the Luxembourgers with respect to languages. The results show that Luxembourgish is widely identified as the language that belongs the most to Luxembourg. There is also a majority which identifies French and German, the other two official languages, as belonging to the country as well. In contrast, Portuguese, the strongest minority language in the country (16% of the total population have Portuguese roots; STATEC, 2019), is not largely attributed to the country. Compared to English, however, for which the answers show a symmetrical distribution (which indicates indecision among the respondents), it belongs more to Luxembourg.

We also find this clear hierarchy of languages present in the country in the aggregated user comments from RTL, as a query of the vector similarities to the country name *Lëtzebuerg* for the same five languages shows:

> *Lëtzebuergesch* ("Luxembourgish", 0.368894), *Franséisch* ("French", 0.296720), *Däitsch* ("German", 0.288161), *Englesch* ("English", 0.276643), *Portugisesch* ("Portuguese", 0.272050)

Remember that the closer a word vector for a language in the model is to the comparison vector, the higher its discursive proximity, that is, its likelihood of appearing in comparable semantic-syntactic contexts, for example, discussions about multilingualism. The query results show that the three-tier hierarchy of languages established in the survey data is also present in the aggregated user comments, with *Lëtzebuergesch* being the closest to *Lëtzebuerg*, followed by *Franséisch* and *Däitsch*, and *Englesch* and *Portugisesch* at a greater distance.

This connection becomes even clearer when asking about the presence of the different languages in everyday life, for example in the public. Traditionally, the majority of public writing is in French and German, but in recent years there has been a substantial increase in Luxembourgish (due to its societal revaluation) and English (as a sign of internationalization).

This aspect of discourse is reflected in the embedding model, for example, in the vector similarities of the variants *Public* ("the public" Romance origin), *Ëffentlechkeet* ("the public" Germanic origin), and Alldag ("everyday life") for the same languages:

TABLE 1 | Belonging of the most important languages to Luxembourg | $N = 1,831$, $p < 0.001$ (χ^2).

"… belongs to Luxembourg"	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
Luxembourgish	91.1	6.5	1.1	1.0	0.2
French	36.6	41.9	8.8	6.2	6.4
German	25.6	47.2	13.2	9.9	4.1
English	9.3	30.0	22.6	27.7	10.5
Portuguese	13.7	35.3	17.2	17.3	16.5

TABLE 2 | Language visibility in public space | $N = 1,831$, $p < 0.001$ (χ^2).

"… should be more visible in public space"	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
Luxembourgish	76.6	16.3	6.2	0.4	0.5
French	1.9	7.2	36.2	28.6	26.0
German	5.8	15.2	40.0	23.0	15.9
English	7.0	18.4	35.2	21.3	18.1
Portuguese	0.6	3.8	22.9	24.1	48.5

- **Alldag**: *Englesch* (0.274927), *Franséisch*, 0.241679), *Lëtzebuergesch* (0.233781), *Däitsch* (0.191266), *Portugisesch* (0.089524)
- **Public**: *Lëtzebuergesch* (0.114520), *Franséisch* (0.081171), *Englesch* (0.048003), *Portugisesch* (0.032752), *Däitsch* (−0.030943)
- **Ëffentlechkeet**: *Englesch* (0.158184), *Lëtzebuergesch* (0.152099), *Franséisch* (0.111701), *Däitsch* (0.044319), *Portugisesch* (−0.0518915)

In all cases, German and Portuguese occupy the lower places, which above all reflects the fact that both languages are hardly discussed in the discourse. In contrast, Luxembourgish and English (on the upswing), together with French (perceived as too strongly present), form the discursive center of the discussion about the languages in the country. If we query specific aspects of written language in public, on the other hand, for example for *Stroosseschëlder* ("street signs"), we get an accurate ranking of the presence of the different languages in the public sphere (see Purschke, 2020 for a quantitative survey of the Luxembourg City linguistic landscape):

- **Stroosseschëlder:** *Franséisch* (0.255751), *Däitsch*, (0.240960), *Lëtzebuergesch* (0.240203), *Englesch* (0.205646), *Portugisesch* (0.187109)

There is a societal demand for a greater presence of Luxembourgish in the public sphere, which is also related to the demographic development of the country, and which is reflected in the survey data in the question of which languages should be more visible in public (**Table 2**). The vast majority of respondents expresses a wish for Luxembourgish to have a greater presence as opposed to the other languages in question. The respondents in particular reject French, which has been assigned a problematic role in the public discussion due to its strong presence among the foreign working population, and Portuguese, which is identified as a language linked to migration in Luxembourg.

The State of Multilingualism

Another section of the questionnaire deals with the assessment of the situation of multilingualism in the country. In this context, we asked the respondents a three-part question that addresses different attitude-related aspects. First the participants had to assess the *current state* of multilingualism. Second, the participants should assess a prognostic statement regarding the future development of multilingualism (*future state*). And third, we used a statement on the preservation of multilingualism in the country to establish the normative dimension (*target state*) of attitudes. By comparing the different answers, we can determine the *attitudinal horizon* of the respondents regarding this complex (**Table 3**).

The results show that Luxembourgers in general have a positive attitude toward multilingualism. A large majority of respondents want it to persist. A majority of the participants also make a positive assessment of the current situation and future development of multilingualism. However, this result also shows that, on the one hand, a substantially larger proportion of the respondents (∼25% each) also see problems in this context, and, on the other hand, the respondents assess the future development of the situation slightly more skeptically than the current state (we make the same observation for similar questions in the study).

A potential reason for the shape of this attitudinal horizon can be found in the comment data. The analysis of the 10 nearest word vectors to the term *Méisproochegkeet* "multilingualism" points to several discursive contexts:

> *Villsproochegkeet* ("multilingualism", 0.738968), *Identitéit* ("identity", 0.654430), *Nationalsprooch* ("*national language*", 0.645417), *Bankeplaz* ("*banking center*", 0.629268),

TABLE 3 | The state of multilingualism | $N = 1,825$, $p < 0.001$ (χ^2).

"Multilingualism in Luxembourg…"	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
Is functioning without problems	16.7	47.6	13.6	16.1	6.0
Will function without problems in the future	15.3	42.1	15.0	20.9	6.7
Should remain	50.1	33.2	7.7	6.0	3.1

TABLE 4 | The status of Luxembourgish $N = 1,829$, $p < 0.001$ (χ^2).

	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
"Luxembourgish is an independent language"	73.9	20.0	3.6	2.1	0.4
"Luxembourgish should be officially recognized as language of the EU"	69.6	15.6	6.2	4.9	3.6
"Newcomers to Luxembourg should learn Luxembourgish"	61.2	31.3	6.6	0.6	0.3

> *Souveränitéit* ("*sovereignity*", 0.627935), *Sprooch* ("*language*", 0.623087), *Orthographie* ("*orthography*", 0.618942), *Ekonomie* ("*economy*", 0.609412), *Zivilisatioun* ("*civilization*", 0.609396), *Économie* ("*economy*" Romance variant, 0.605071)

First, we see a close relationship with other language-related concepts, which can be expected due to the model logic of word embedding. Second, and more interestingly, multilingualism appears in a discursive context that deals with societal and national issues (*Identitéit*, *Souveränitéit*). Against the backdrop of the public discourse on the language situation in recent years, this shows above all the close connection between language- and identity-related questions that partly shape the public discussion in Luxembourg, especially in political and right-wing populist contexts. Third, the word vectors that refer to economic aspects (*Ekonomie*, *Bankeplaz*) demonstrate the close interdependence of the Luxembourg economic model with multilingualism: the private sector and the financial industry mostly employ foreign workers. The increase in this population group through migration and cross-border commuting, as well as the associated presence of languages other than Luxembourgish in public, are the rated breaking points in the societal discussion on multilingualism.

The Status of Luxembourgish

Another central issue in the public discussion concerns the role of Luxembourgish, that is, its status as a language. Linguistically speaking, Luxembourgish is a Moselle-Franconian dialect and is therefore closely related to the German regional languages (Gilles, 2019). Despite the fact that Luxembourgish has been declared the national language by law in 1984—and thus has an official language status—there are still stances in the discourse that describe Luxembourgish as a German dialect (as opposed to German and French as fully-developed and prestigious languages of culture; Sieburg and Weimann, 2014). However, if we ask the participants about the status of Luxembourgish, a large majority confirm its official status as a language (**Table 4**). At the same time, 20% of the respondents only somewhat agree to the question. This assessment coincides with further judgments about the status of Luxembourgish in the data: an equally large majority supports the recognition of the language at EU level. In addition, there is a clear positioning (and expectation of linguistic integration) vis-à-vis immigrants with regard to language acquisition (remember the stance in *Example 1*).

Contrasting the respondents' attitudinal horizon regarding Luxembourgish with the public stances in the comment data also reveals a correspondence. In the aggregated data there is a greater discursive proximity from *Lëtzebuergesch* to the vector for *Sprooch* ("language," 0.642606) than to the vector for *Dialekt* ("dialect," 0.487487). This means that Luxembourgish is discussed more likely in the context of a (national) language than its origin as a dialect of German.

A characteristic (and strength) of Luxembourgish is its high degree of linguistic plasticity. The language has a high proportion of elements of German or French origin and continues to integrate them without problems. In the current discourse climate, however, this flexibility is sometimes seen as problematic, for example by language activists who are committed to keeping Luxembourgish "clean" from "foreign" influences. A good indicator question in the questionnaire for this connection is that of the assumed linguistic influences on Luxembourgish in the future (**Table 5**). As can be seen, the respondents see a growing influence of English and French on the language, not so much of German. Interestingly, this assessment somewhat contradicts linguistic reality. As Conrad (2017) shows, younger speakers in particular show a clear preference for the Germanic variants when choosing between parallel phonological variants, not toward the Romance variants. In this respect, we can read the result as an assessment of the *assumed cultural influence* of the languages in Luxembourg society rather than of their *factual linguistic influence* on Luxembourgish.

Again, we can see the same assessment in the comment data. Querying for the 20 nearest neighboring vectors for different combinations of *Lëtzeguergesch* + x (i.e., *Afloss* "influence," *Entwécklung* "development," *Zukunft* "future"), it becomes apparent that French and English are always in a greater discursive proximity than German:

TABLE 5 | Future influences on Luxembourgish | $N = 1{,}826$, $p < 0.001$ (χ^2).

"The influence of … on Luxembourgish will grow in the future"	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
German	4.0	22.7	34.3	31.3	7.7
English	8.2	42.2	20.8	20.1	8.7
French	11.0	41.1	26.8	16.2	4.9

- **Afloss**: ***Franséisch*** ("French"), ***Englesch*** ("English"), ***Franzéisch*** ("French," spelling variant), *Impakt* ("impact"), *Zougrëff* ("access"), **franséisch** ("French," ADJ/lower case N), ***Franséich*** ("French," spelling variant), *Accent* ("accent," Romance variant), *Akzent* ("accent," Germanic variant), ***englesch*** ("English," ADJ/lower case N), ***Däitsch*** ("German"), *Letzebuergech* ("Luxembourgish," spelling variant), *Lëtzbuergesch* ("Luxembourgish," spelling variant), Lëtzebuergescht ("Luxembourgish," inflection form), *Urecht* ("entitlement to")
- **Entwécklung**: *Sprooch* ("language"), *Orthographie* ("orthography"), *Integratioun* ("integration"), ***Franséisch*** ("French"), *Schreifweis* ("spelling"), ***Englesch*** ("English"), *Grammatik* ("grammar"), *Allgemengbildung* ("general education"), *Literatur* ("literature"), *Rechtschreiwung* ("orthography"), ***Franséischt*** ("French," inflection form), *Lëtzebuergescht* ("Luxembourgish," inflection form), *Kommunikatioun* ("communication"), *Evolutioun* ("evolution"), *Mammesprooch* ("mother tongue")
- **Zukunft**: ***Franséisch*** ("French"), *Sprooch* ("language"), ***Englesch*** ("English"), *Mammesprooch* ("mother tongue), ***Sprooche*** ("languages," n-rule form), *franséisch* ("French," ADJ/lower case N), ***Lëtzebuergescht*** ("Luxembourgish," inflection form), *sprooch* ("talk," imperative/"language," lower case N), ***englesch*** ("English," ADJ/lower case N), *Integratioun* ("integration"), *Landessprooch* ("national language"), ***Franzéisch*** ("French," spelling variant), ***Franséich*** ("French," spelling variant), *Sproochen* ("languages"), ***Franséischt*** ("French," inflection form)

Apart from the fact that in a word embedding model the different language names are inevitably close to each other (due to concept similarity), the different sequences and constellations indicate similar prognostic evaluations regarding the development of Luxembourgish. Ultimately, these constellations in the discourse mirror assumptions about the *global cultural dynamics* of the country (demographically and economically), and the languages are representative of this.

The Language-Identity Link

The comment data in particular reveal a close connection between linguistic concepts and those that belong more in the area of identity and nationality. For the 30 closest neighbors to the word vector *Sprooch* "language," the embedding model shows elements that we can link to different semantic domains (**Table 6**).

In addition to the language names for French and Luxembourgish (not German, though!), there are a number of other related concepts that we can assign to the linguistic context of the term *Sprooch*, including *Grammaire/Grammatik* "grammar," W*eltsprooch* "world language," or spelling variants and inflection forms of the concept. However, there are also a number of concepts that place the word in other semantic domains, namely references to words that relate to political and nation-state contexts, and words that relate to individual or collective identity constructions. This discursive proximity of different semantic domains also indicates the range of possible discursive contexts in which the concept of language appears in the comment data. In this context, we can read the identity- and nation-related concepts as an indication of the close connection of language, identity, and nation in the discourse, which is in fact a characteristic of the public discussion about language in recent years. Garcia (2014) diagnoses a strong politicization and ideological charging of the language discussion in Luxembourg. In this context, it is also revealing to observe that many Luxembourgers, when referencing Luxembourgish, use the term *eis Sprooch* "our language" (see above, *Example 2*), that is, they directly identify the language with the political community—as opposed to the other official languages of the country, French and German.

TABLE 6 | Semantic domains of nearest neighbors to *sprooch* "language."

Linguistic concepts	*Sprooch* ("talk," imperative/"language," lower case N), *Schreifweis* ("spelling"), *Sproch* ("language," spelling variant, "saying"), *Orthographie* ("orthography"), *Friemsprooch* ("foreign language"), *Sproochen* ("languages"), *Sprooche* ("languages," n-rule form), *Grammatik* ("grammar," Germanic variant), *Grammaire* ("grammar," Romance variant), *Méisproochegkeet* ("multilingualism"), *Rechtschreiwung* ("orthography"), *Villsproochegkeet* ("multilingualism"), *Weltsprooch* ("world language"), *Ëmgangssprooch* ("colloquial language"), *Mondart* ("dialect"), *Iwwersetzung* ("translation")
Language concepts	*Lëtzebuergesch* ("Luxembourgish"), *Franséisch* ("French"), *Lëtzebuergescht"* "Luxembourgish," inflection form)
National concepts	*Landessprooch* ("national language"), *Nationalsprooch* ("national language"), *Nationalitéit* ("nationality"), *Amtssprooch* ("official language"), *Gesetzgebung* ("legislation"), *Nationalhymne* ("national anthem"), *Verfassung* ("constitution"), *Integratioun* ("integration")
Identity concepts	*Mammesprooch* ("mother tongue"), *Identitéit* ("identity"), *Kultur* ("culture")

We find an additional illustration of this nexus by querying the vector similarities for the concepts *Mammesprooch* "mother tongue" and *Friemsprooch* "foreign language" with the vectors for the most important languages:

- ***Friemsprooch***: *Franséisch* (0.562637), *Englesch* (0.552961), *Portugisesch* (0.540379), *Däitsch* (0.516047), *Lëtzebuergesch* (0.507465)
- ***Mammesprooch***: *Lëtzebuergesch* (0.553955), *Franséisch* (0.547082), *Däitsch* (0.517634), *Portugisesch* (0.512621), *Englesch* (0.510809)

As we can see, the contextual similarity is different for the two concepts, with Luxembourgish being closest to the concept mother tongue and furthest away from the concept foreign language, unlike English. German and Portuguese occupy middle positions in both queries. A possible reason for this could again be the fact that these languages are not assigned a problematic role for the organization of multilingualism in the current discourse. Most interestingly, French is close to both of the concepts queried, reflecting its overall prominent role in the discourse: the language is seen as both "foreign" (linked to work-related migration) and "native" (historically rooted in Luxembourg multilingualism).

Language Preferences in Everyday Practice

The close connection between language and self-image is not only evident in the discussions about language, but also in everyday preferences for certain languages. We asked a number of questions in the questionnaire that not only provide information about specific language preferences, but also demonstrate that the language regime in Luxembourg is currently on the move. For example, the participants were asked which languages are important to them in everyday life (**Table 7**).

As the data show, there is a clear hierarchization of the different languages in terms of their practical use in everyday life, with Luxembourgish being by far the most important tool in practice. This statement also partially reflects the composition of the sample: the majority of the study participants are native Luxembourgers with Luxembourgish as (one of) their mother tongue(s). In addition, the data also confirm the important role of French in Luxembourg multilingualism. More interesting than the general usefulness are therefore the questions about the specific language preferences in everyday situations, for example, when watching TV news (**Table 8**).

On the one hand, it becomes clear that the respondents do in fact have a strong preference for Luxembourgish (1st choice), but there is also an effect of the domain specificity of Luxembourgish multilingualism: In practice, many Luxembourgers mainly watch German television (2nd choice), partly because of the linguistic proximity to Luxembourgish, but also because the number of Luxembourgish channels is limited (to RTL). On the other hand, the 3rd choice is particularly interesting, in which the test subjects mostly choose between English and French. While the summary result seems to prefer French as 3rd choice, a look at the answers of the different age groups (**Table 9**) shows that the preference shifts from French to English with decreasing age.

We can compare these preferences with the RTL authors' language choices in the comment data, since writing a comment online also represents a (media-related) everyday situation. However, since writing in Luxembourgish is still a challenge for many Luxembourgers, this situation is far less routinized than watching TV news. On the other hand, the choice of language is influenced in part by the larger communicative context of the platform with Luxembourgish as default language for both news texts and comments. Based on the automatic language detection and considering only texts with more than 200 characters (see *Footnote 1* for information on detection accuracy), we find that the vast majority of texts is written in Luxembourgish (343,336 of 357,163 texts total), as opposed to 10,268 texts in German, 2,915 in French, and 399 in English—the remaining texts are mostly wrongly identified Luxembourgish texts labeled as Dutch. This result proves that—at least on the RTL platform—Luxembourgish has established itself as the default written language, but it also shows that German is preferred over French as an alternative language.

TABLE 7 | General language preference in everyday life | $N = 1,824, p < 0.001(\chi^2)$.

"… is an important tool for me in everyday life"	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
Luxembourgish	87.0	10.7	1.0	1.0	0.3
French	35.1	42.9	9.0	6.8	6.2
German	18.5	31.9	18.1	23.5	8.0
English	13.7	24.0	19.7	27.9	14.8
Portuguese	1.5	5.1	7.5	21.2	64.7

TABLE 8 | Language preference when watching TV news | $N = 1{,}827$, $p < 0.001$ (χ^2).

"Which language do you prefer when watching the news on TV?"	Luxembourgish (%)	German (%)	French (%)	English (%)	Portuguese (%)	Italian (%)
1st choice	68.3	23.4	4.6	3.3	0.4	0.0
2nd choice	17.9	61.7	11.6	7.7	0.8	0.2
3rd choice	8.6	11.2	46.8	31.7	1.0	0.6

TABLE 9 | Language preference TV news, 3rd choice by AGE | $N = 1{,}827$, $p < 0.001$ (χ^2).

"Which language do you prefer when watching the news on TV?"	Luxembourgish (%)	German (%)	French (%)	English (%)	Portuguese (%)	Italian (%)
≤24	9.3	10.0	36.1	41.1	2.4	0.7
25–34	5.5	11.2	45.4	36.6	1.1	0.2
35–44	8.3	8.9	50.5	31.5	0.6	0.3
45–54	9.3	13.4	56.0	19.4	0.0	1.4
55–64	10.3	11.8	54.9	21.5	0.5	1.3
≥65	20.5	17.9	47.4	14.1	0.0	0.0

More generally speaking, and in line with most processes of language change, the age of the speakers is a determining factor for their linguistic orientation in everyday life—and thus for attitudes toward Luxembourg multilingualism. In the questionnaire data, age is the main demographic structuring factor explaining differences in attitudes. We can assume that the language regime will shift substantially in favor of English in the next few years, especially through the shift in the linguistic preferences of the young speakers—but also in view of the continuing internationalization of the resident population. In 2019, there was even a public petition to establish English as an official language in administrative contexts next to French and German[5]. In view of the many languages and sociocultural factors involved in this dynamic, it is hardly possible, though, to make a forecast about the development of the language regime as a whole.

DISCUSSION

Following the analysis, we discuss some methodological aspects in more detail below. This concerns the reconstruction of attitudes with the help of word embedding models as well as the collection of language attitudes data using crowdsourcing, but also the automatic orthographic normalization of Luxembourgish texts and potential limitations of the overall approach.

Reconstructing Attitudes Using Representation Learning

The comparative analysis of attitudes toward multilingualism in Luxemburg has shown that word embedding models can be successfully used for the reconstruction of attitudes in free text data. The quantitative modeling brings to light discursive attitudinal patterns that represent the sum of many individual stances, without each individual stance itself necessarily being a direct expression of the aggregated attitude. During the preprocessing of the data, however, we have seen that and to what extent word embedding models are susceptible to the selection of the hyperparameters for training, that is, the number of vector dimensions or the window length for word contextualization (Goldberg, 2017; Pierrejean and Tanguy, 2018). The same holds true for data-intrinsic factors like the total number of words, vocabulary size, and word frequency range. Depending on the setting of the hyperparameters, different training results can be expected, especially in the upper and lower frequency range of the vocabulary.

In this respect, the orthographic normalization of the texts before training the data has a clear impact on the word embedding model on which the analysis is based. However, the comparison of different model solutions shows that the vector space is relatively stable for the concepts discussed in the present study, since it is usually a matter of words in the middle range of the frequency spectrum. For example, the 10 nearest-neighbor vectors for the word *Sprooch* "language" largely match before and after the orthographic correction:

- **Before normalization:** *Sproch* ("saying, language," spelling variant, 0.842746), *Mammesprooch* ("mother tongue," 0.800106), *Landessprooch* ("national language," 0.769282), *Schreifweis* ("spelling," 0.711668), *Nationalsprooch* ("national language," 0.709543), *Identitéit* ("identity," 0.700674), *Mammesproch* ("mother tongue," spelling variant, 0.696856), *Orthographie* ("orthography," 0.691917), *Mammensprooch* ("mother tongue," spelling variant, 0.681196), *Sproochen* ("languages," 0.673093)
- **After normalization:** *Mammesprooch* ("mother tongue," 0.814756), *sprooch* ("talk," imperative/"language," lower case N, 0.771097), *Landessprooch* ("national language," 0.759516), *Schreifweis* ("spelling," 0.751803), *Sproch* ("saying, language," spelling variant, 0.723642), *Nationalsprooch* ("national language," 0.723429), *Orthographie* ("orthography," 0.701390), *Identitéit* ("identity," 0.692551), *Friemsprooch* ("foreign language," 0.660245), *Nationalitéit* ("nationality," 0.656720)

While the nearest neighbors represent more or less the same concepts, the example also demonstrates the value of orthographic normalization. After the correction process, several spelling variants are no longer among the nearest neighbors (and no longer in the vocabulary of the model). Nevertheless, orthographic normalization brings with it some methodological and practical challenges, for example, the lack of distinction between *Sproch* as a common spelling variant of *Sprooch* "language" and as a separate lemma with the meaning "saying."

[5] See https://chd.lu, public petition No. 1404, for further details. In Luxembourg, any resident can submit a public petition on the website of the parliament. Provided it gets enough support (the current threshold is at 4,500 signatures), it will be officially discussed in parliament.

Orthographic Normalization

Given the diverse sources of orthographic variation in Luxembourgish, the normalization of the texts is an important step in preparing the data for analysis. Normalization (using the current build of the *spellux* package) reduces the number of unique words in the data set and ensures more consistent vector representations by integrating orthographic variants into the basic lemma. The pipeline developed for processing the data works reliably, but the correction does not produce error-free texts. On the one hand, this is due to the number of orthographic variants that are not yet captured by the correction resources. On the other hand, the correction routine also produces a number of *false positives* and *false negatives*: Some words that can be identified in context as misspellings of lemmas also exist as an independent lemma with a different meaning (remember the example *Sprooch—Sproch*). In this case, we do not correct the word, due to a false-positive validation of the word form in the lemma list. At the same time, in the course of normalization, we do correct a number of word forms that represent misspellings to lemmas that are either contextually incorrect (because the word form is listed as a variant in the correction dictionary) or wrongly evaluated as a correction candidate in the comparison with the correction resources. As for the peculiarities of the writing system (n-rule), the *spellux* package has a dedicated rule-based correction routine for this context rule. Given the large amount of exceptions from the base rule (e.g., for personal and country names), however, we still cannot capture all cases when automatically correcting texts. We must therefore establish criteria for orthographic normalization to evaluate the advantages and disadvantages of an automated text correction, also in light of its impact on model training.

The comparison of the corrections made to an example text is helpful for illustration of the effects and challenges of automatic normalization. Misspellings in the original text are marked in *italics* (including n-rule errors). Correct corrections in the normalized text are marked in **bold**, incorrect corrections are underlined and variants that have not been corrected remain in *italics*.

> **Before Normalization:**
> *Den* Grand-Duc huet gerad *eso* Recht fir no sengem *Gewessen* ze entscheeden, *an* wann dat *den sogenannte* Spëtzepolitiker *an verschiedene* Journalisten net *gefaellt* dann *haet schons laengst* versicht solle *gin* Verfassung *dementsprechend* ze *aenneren*. *Den* Problem do, huet eng *Kéer missen kommen.An dann welle* Politiker an *eso* engem Dossier *wei* Liewen an den Doud *Haptwuert huen*, mat engem *débat doriwer wo* sie den Niveau *emmer mee* erof *zéen an* Leit um *terrain kennen* herno kucken dass se kloer kommen
> **After Normalization:**
> **De** Grand-Duc huet grad **esou** Recht fir no sengem **Gewëssen** ze entscheeden, **a** wann dat **de sougenannte** Spëtzepolitiker **a verschidde** Journalisten net **gefält** dann **hätt schonns längst** versicht solle **gi** Verfassung **deementspriechend** ze **änneren**. **De** Problem do, huet eng *Kéer* **misse kommen. An da wëlle** Politiker an **esou** engem Dossier *wei* Liewen an den Doud *Haptwuert* **hunn**, mat engem *débat* **doriwwer** *wo* Sie den Niveau **ëmmer** *mee* erof **zéien a** Leit um **Terrain** *kennen* herno kucken dass se kloer kommen.

As we can see, the automatic correction replaces most of the incorrect spellings with the correct ones. In addition, there are also some false corrections, e.g., *sie*$_{[\text{before}]}$ ("they") is corrected to *Sie*$_{[\text{after}]}$ ("B," musical note, plural + n-rule reduction) instead of the correct pronoun spelling *si*. No correction was made to some variants, be it because no variant–lemma pair was found in the correction resources (*Kéer* for *Kéier* "time, occasion"), be it because the variant matches with the wrong lemma in the lemma list (*mee*, meaning *méi* "more" in this context, matches with the lemma *mee* "but"). For these cases, we must expand the correction dictionary with additional spelling variants and finetune it. A final type of change relates to the form *kommen.an* in the original text. This is an artifact of tokenization and is detected during the correction routine. Regardless of such problems, the current correction architecture can already substantially consolidate the vocabulary of the data set.

A number of factors must be taken into account for further developing the *spellux* package:

- We must expand the correction dictionary to include more spelling variants that are present in the data but have not been recorded so far to reduce the number of unidentifiable variants.
- We must evaluate the use of case-sensitive models for correction and training: while the current workflow increases the number of remaining spelling variants in the corpus (e.g., *Lëtzebuergesch* N vs. *lëtzebuergesch* ADJ/lower-case N), using a lower-case model would produce a higher number of homographic lemmas and therefore reduce correction accuracy.
- We should integrate additional contextual cues to word disambiguation in order to determine correction candidates for variants without corresponding lemma in the existing correction resources. This includes candidate evaluation based on POS tags as well as on n-grams.
- We should systematically evaluate the training parameters for the correction resources with regard to their impact on correction performance. This applies above all to the correction frequency threshold for the spelling variants when building the correction dictionary, but also to the minimum frequency threshold for words when training the correction model for the entire data set, and to the similarity threshold for candidate evaluation in the correction workflow.
- We must consider lemmatization of words to further consolidate the vocabulary as well as removing stop words. Both the *spellux* package and the language support for Luxembourgish in *spaCy* have inbuilt options for lemmatization and stop word removal. The content analysis, however, shows that in some cases stop words (remember the example *eis Sprooch* "our language") are part of discursive patterns that can be meaningfully interpreted.

Measuring Attitudes Quantitatively Using a Mobile Crowdsourcing App

In the Schnëssen app, we use a classical questionnaire survey for data collection, in which the answers of the respondents are quantified using scaling. Compared to qualitative studies that work with interviews or ethnographic methods, this approach has the advantage of an easier evaluation and generalizability of the data. Results do not have to be condensed qualitatively based on categories derived from the data. Conversely, quantitative methods are not suitable for all aspects of attitudes research (see Casper, 2002 for a discussion), especially assuming that attitudes are situated evaluation routines that arise and come into play in practice (Purschke, 2015). For example, the complex Luxembourg multilingualism is not only organized according to social domains, which are relatively easy to query in a questionnaire study. In addition, the daily organization of language practice is highly dependent on individual factors, for example, the language skills of interlocuters, the social environment, and everyday routines, that influence the language preferences and the situational choice of a language. These can hardly be recorded using a general quantitative questionnaire.

Nevertheless, there are societal macro-conditions that lead to many people having comparable experiences that are anchored in their everyday social practice. This concerns, for example, language teaching in schools, which is partly responsible for the current poor image of French in the country, since the language is taught in a very formal and norm-oriented manner. The same applies to the country's global socio-economic demographic development that affects the language regime as a whole and that is being negotiated in public discourse, as can be seen from the RTL comments. Therefore, the questions in the questionnaire focus primarily on such aspects. In this way, we can ensure that the respondents already have the attitudes to be surveyed at their disposal because they are part of their everyday life experience.

The type of data collection using crowdsourcing also plays an important role in the composition and analysis of the data (see Entringer et al., forthcoming for a discussion). In principle, app-based crowdsourcing of linguistic data enables the collection of a large data set with comparatively little effort. However, we have to invest a lot of work in social media activities and public outreach in order to acquire enough respondents and to motivate them to a continued participation in longer survey campaigns. One technical challenge of the data set stems from the difficulties with matching social profiles and questionnaires. As a result, some of the completed questionnaires could not be considered for the analysis. However, on the basis of this identification, we can also compare the results of the questionnaire with the actual language use of the same participants in the app's recording task, for example, with regard to their attitudes toward German and French and their individual choice between competing lexical or phonological variants that originate from German or French. With regard to the demographic bias of the data basis, a targeted expansion of the sample by foreign residents and cross-border commuters would be desirable to get a more differentiated and comprehensive view of existing attitudes. To do this, we must also consider translating the questionnaire into other languages.

Limitations of the Approach and Implications for Attitudes Research

The comparison of results using complementary data sets has proven to be insightful. For many questions from the questionnaire, we find corroborating evidence in the aggregated comment data. However, this this does not apply to all contexts. To illustrate this, we use one last question complex asking the participants about their attitudinal horizon for writing Luxembourgish (**Table 10**).

The first question is an example that can be easily substantiated with the comment data even without querying the model. A large majority of respondents say that they write texts in Luxembourgish in everyday life, and this is exactly what the authors of the comments on RTL.lu do. The second question, on the other hand, cannot be easily converted into an informative query: the combination of s*chreiwen* ("to write") and *Zukunft* ("future") yields exclusively related verb concepts; the combination of *schreiwen*, *Zukunft*, and *Lëtzebuergesch* results mostly in related language concepts. Additionally, the third question documents potential discrepancies between the two data sets. While the majority of those questioned in the Schnëssen survey express a normative orientation toward the official spelling rules, the extent of orthographic variation in the comments proves the lack of practical implementation of these spelling rules. In view of the ongoing standardization of Luxembourgish, we can assume that the attitudinal orientation toward the norm precedes the actual practical acquisition of writing skills.

For the contrastive study of language attitudes, these findings mean that extensive contextual knowledge of the sociocultural, linguistic, and language-political context may be necessary to relate the results of the different analyses to one another in a meaningful way. At the same time, we can use this approach to investigate attitudes comprehensively (i.e., through complementary evidence from different datasets) and differentiated (e.g., regarding the difference between stances in discourse and connected underlying attitudes). Taken together, the results open up interesting perspectives both for attitudes research and for a culturally aware computational processing of

TABLE 10 | Writing practice in Luxembourgish | $N = 1,828$, $p < 0.001$ (χ^2).

	Agree (%)	Somewhat yes (%)	Neither nor (%)	Somewhat no (%)	Disagree (%)
"I do write texts in Luxembourgish in everyday life"	72.9	19.1	2.1	5.0	0.9
"I will write more texts in Luxembourgish in the future"	40.7	19.8	31.2	6.5	1.9
"When writing Luxembourgish, I should stick to the official rules"	37.9	41.3	10.8	8.1	1.9

text data. One particular challenge for further research in this context is the direct implementation of quantitative attitudes data in the training of word embedding models as a form of *social retrofitting* of such models.

CONCLUSION

The aim of the present study was the contrasting investigation of language attitudes using the example of free text data from user comments and quantitative attitudes data from a survey. We have shown that sociolinguistic and computational methods can be successfully combined for the analysis of societal issues. This is confirmed by the correspondences between the attitudes reconstructed from the aggregated text data and the attitudes surveyed with the questionnaire. The results testify to the differentiated attitudinal horizons of the Luxembourgers concerning multilingualism in general and the individual languages in the language regime. The study also demonstrates the potential of computational sociolinguistics, at the center of which is the analysis of language as a sociocultural phenomenon. However, the work with the different approaches and data types also shows that we cannot interpret the results of the analysis without contextual knowledge about the sociolinguistic situation and the structure and dynamics of public discourse. Only the comparative analysis and embedding of the results in the larger sociocultural context allows us to make reliable statements about the research question at hand. It has also become clear that computational sociolinguistics needs a solid linguistic-theoretical basis and standardized technical-methodological procedures in order to fully unfold its potential for the study of language as a cultural phenomenon.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found on Zenodo: Luxembourgish word embedding model (user comments from RTL.lu): doi: 10.5281/zenodo.3978066; Schnëssen attitudes survey data: doi: 10.5281/zenodo.3978084.

ETHICS STATEMENT

This research is in line with the rules and regulations for research ethics at the University of Luxembourg as stated in the official Ethics Review Committee policy (adopted by the Board of Governors at its meeting of October 25, 2019). The survey data from the Schnëssen project were collected on the basis of informed consent and were strictly anonymized for storage, processing, and analysis. The text data from the RTL news platform were provided by RTL in anonymous form. Identification of individuals based on the available data was not possible at any time.

AUTHOR CONTRIBUTIONS

All contributions (analyses, code, text) were made by CP. The data sources for the analyses were developed, collected, and prepared in collaboration with the colleagues from the projects Schnëssen and STRIPS.

FUNDING

The projects STRIPS and Schnëssen were internally funded by the University of Luxembourg.

ACKNOWLEDGMENTS

The author thanks the RTL Media Group Luxembourg for providing the text data used for the analysis. Furthermore, the author thanks the participants of the Schnëssen survey for their contributions. Many thanks also go to the two reviewers for their constructive and helpful advice.

REFERENCES

Abu-Jbara, A., Hassan, A., and Radev, D. (2012). "AttitudeMiner: mining attitude from online discussions," in *Proceedings of the NAACL-HLT 2012: Demonstration Session* (Montréal, QC), 33–36. Available online at: https://www.aclweb.org/anthology/N12-3009

Albarracín, D., and Johnson, B. T. (2018). *The Handbook of Attitudes,* 2nd Edn. New York, NY: Routledge.

Aman, S., and Szpakowicz, S. (2008). "Using Roget's thesaurus for fine-grained emotion recognition," in *Proceedings of the IJCNLP 2008*, 296–302. Available online at: https://www.aclweb.org/anthology/I08-1041

Behrend, T. S., Sharek, D. J., Meade, A. W., and Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behav. Res.* 43:800. doi: 10.3758/s13428-011-0081-0

Bellamy, J., and Horner, K. (2018). Ein Mischmasch aus Deutsch und Französisch: ideological tensions in young people's discursive constructions of luxembourgish. *Sociolinguist. Stud.* 12, 323–342. doi: 10.1558/sols.34809

Belling, L., and de Bres, J. (2014). Digital superdiversity in Luxembourg: the role of Luxembourgish in a multilingual Facebook group. *Discourse Context Media* 4–5, 74–86. doi: 10.1016/j.dcm.2014.03.002

Bender, E., and Koller, A. (2020). "Climbing towards NLU: on meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the ACL,* 5185–5198. doi: 10.18653/v1/2020.acl-main.463

Boucouvalas, A. C. (2003). "Real time text-to-emotion engine for expressive internet communications", in *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*, eds G. Riva, F. Davide, and W.A. IJsselsteijn (Amsterdam: Ios Press), 306–318.

Broadwell, G., Stromer-Galley, J., Strzalkowski, T., Shaikh, S., Taylor, S., Boz, U., et al. (2013). Modeling socio-cultural phenomena in discourse. *J. Nat. Lang. Eng.* 19, 213–257. doi: 10.1017/S1351324911000386

Casper, K. (2002). *Spracheinstellungen. Theorie und Messung.* Heidelberg: Books on Demand.

Chaumartin, F.-R. (2007). "UPAR7: a knowledge-based system for headline sentiment tagging," in *Proceedings of the 4th International Workshop on Semantic Evaluations* (Prague), 422–425. Available online at: https://www.aclweb.org/anthology/S07-1094

Chuang, J.-H., and Hsieh, S.-K. (2015). Stance classification on PTT comments. *PACLIC* 29, 27–36.

Chuang, Z.-J., and Wu, C.-H. (2004). Multimodal emotion recognition from speech and text. *Comput. Linguist. Chin. Lang. Process* 9, 45–62. Available online at: https://www.aclweb.org/anthology/O04-3004

Conrad, F. (2017). *Variation durch Sprachkontakt*. Frankfurt am Main; New York, NY: Peter Lang.

Dasigi, P., Guo, W., and Diab, M. (2012). "Genre independent subgroup detection in online discussion threads: a pilot study of implicit attitude using latent textual semantics," in *Proceedings of the 50th Annual Meeting of the ACL*, 65–69. Available online at: https://www.aclweb.org/anthology/P12-2013

De Bres, J. (2014). Competing language ideologies about societal multilingualism among cross-border workers in Luxembourg. *Int. J. Sociol. Lang.* 227, 119–137. doi: 10.1515/ijsl-2013-0091

De Bres, J., and Franziskus, A. (2019). "Language ideologies in conflict at the workplace", in *The Routledge Handbook of Language in Conflict*, eds M. Evans, L. Jeffries, and J. O'Driscoll (London: Routledge), 433–447.

Dong, M., Jurgens, D., Banea, C., and Mihalcea, R. (2019). Perceptions of social roles across cultures. *SocInfo* 2019, 157–172. doi: 10.1007/978-3-030-34971-4_11

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *J. Sociolinguist.* 19, 161–188. doi: 10.1111/josl.12119

Entringer, N., Gilles, P., Martin, S., and Purschke, C. (forthcoming). "Schnëssen. Surveying language dynamics in Luxembourgish with a mobile research app", in *Linguist Vanguard, Special volume "Using Smartphones to Collect Linguistic Data"*, eds A. Leemann and N. Hilton.

Erhart, S., and Fehlen, F. (2011). "Luxembourgish: a success story? A small national language in a multilingual country," in *Handbook of Language and Ethnic Identity*, eds J. A. Fishman and O. Garcia (Oxford: Oxford University Press), 285–298.

Fehlen, F. (2009). *BaleineBis: Une enquête sur un marché linguistique multilingue en profonde mutation. Luxemburgs Sprachenmarkt im Wandel*. Luxembourg City: SESOPI Centre Intercommaunitaire.

Fehlen, F., and Heinz, A. (2016). *Die Luxemburger Mehrsprachigkeit. Ergebnisse einer Volkszählung* Bielefeld: transcript.

Franziskus, A. (2013). *Getting by in a multilingual workplace: the language practices, ideologies and norms of cross-border workers in Luxembourg*. [Ph.D. dissertation]. University of Luxembourg, Luxembourg City, Luxembourg.

Garcia, N. (2014). The paradox of contemporary linguistic nationalism: the case of Luxembourg. *Nations National.* 20, 113–132. doi: 10.1111/nana.12043

Garg, N., Schiebinger, L., Jurafsky, D., and Zoue, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3635–E3644. doi: 10.1073/pnas.1720347115

Garimella, A., Banner, C., Hovy, D., and Mihalcea, R. (2019). "Women's syntactic resilience and men's grammatical luck: gender-bias in part-of-speech tagging and dependency parsing," in *Proceedings of the 57th Annual Meeting of the ACL* (Florence), 3493–3498. doi: 10.18653/v1/P19-1339

Gierschek, D., Gilles, P., Purschke, C., Schommer, C., and Sirajzade, J. (2019). A temporal warehouse for modern luxembourgish text collections. In: *DHBeNeLux* (Liége). Available online at: http://hdl.handle.net/10993/41840

Giles, H., and Marlow, M. (2011). "Theorizing language attitudes: past frameworks, an integrative model, and new directions," in *Annals of the International Communication Association 35*, ed C. Salmon (Thousand Oaks, CA: Sage), 161–197. doi: 10.1080/23808985.2011.11679116

Gilles, P. (2019). "Komplexe Überdachung II: luxemburg. Die genese einer neuen nationalsprache," in *Sprache und Raum. Ein internationales Handbuch der Sprachvariation, Bd. 4: Deutsch*, eds J. Herrgen and J.E. Schmidt (Berlin; Boston: De Gruyter), 1039–1060.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan & Claypool.

Grondelaers, S., and Speelmann, D. (2015). "A quantitative analysis of qualitative free response data. Paradox or new paradigm?" in *Change of Paradigms – New Paradoxes: Recontextualizing Language and Linguistics,* ed J. Daems *(Berlin; Boston: De Gruyter Mouton),* 361–384. doi: 10.1515/9783110435597-021

Hassan, A., Qazvinian, V., and Radev, D. (2010). "What's with the attitude? Identifying sentences with attitude in online discussions," in *Proceedings of the 2010 Conference on EMNLP* (Cambridge, MA), 1245–1255. Available online at: https://www.aclweb.org/anthology/D10-1121

Heblich, S., Lameli, A., and Riener, G. (2015). The impact of regional accents on economic behavior: a lab experiment on linguistic performance, cognitive ratings and economic decisions. *PLoS ONE* 10. doi: 10.1371/journal.pone.0113475

Horner, K., and Weber, J.-J. (2008). The language situation in Luxembourg. *Curr. Issues Lang. Plann.* 9, 69–128. doi: 10.2167/cilp130.0

Hovy, D. (2018). "The social and the neural network: how to make natural language processing about people again," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (New Orleans, LA), 42–49. doi: 10.18653/v1/W18-1106

Jaffe, A. (2013). "Introduction," in *Stance: Sociolinguistic Perspectives*, ed A. Jaffe (Oxford: Oxford University Press), 1–28.

Kloss, H. (1952). *Die Entwicklung neuer germanischer Kultursprachen von 1800 bis 1950*. München: Pohl & Co. Verlagsbuchhandlung.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Soc. Rev.* 84, 905–949. doi: 10.1177/0003122419877135

Kristiansen, T. (2009). The macro-level social meanings of late-modern Danish accents. *Acta Linguist. Hafniensia* 41, 167–192. doi: 10.1080/03740460903364219

Lameli, A., Nitsch, V., Südekum, J., and Wolf, N. (2015). Same same but different: dialects and trade. *German Econ. Rev.* 16, 290–306. doi: 10.1111/geer.12047

Lin, C.-S., Shaikh, S., Stromer-Galley, J., Crowley, J., Strzalkowski, T., and Ravishankar, V. (2013). "Topical positioning: a new method for predicting opinion changes in conversation," in *Proceedings of the Workshop on Language in Social Media 2013* (Atlanta), 41–48. Available online at: https://www.aclweb.org/anthology/W13-1105

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Adv. Neural Inform. Process. Syst.* 26, 3111–3119. Available online at: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.* 5, 101–111. doi: 10.1109/TAFFC.2014.2317187

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). "SentiFul: generating a reliable lexicon for sentiment analysis," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (Amsterdam), 363–368. doi: 10.1109/ACII.2009.5349575

Nguyen, D. (2017). *Text as Social and Cultural Data: A Computational Perspective on Variation in Text* (Ph.D. thesis). Universiteit Twente, Enschede.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Ret.* 2, 1–135. doi: 10.1561/1500000011

Pharao, N., and Kristiansen, T. (2019): Reflections on the relation between direct/indirect methods and explicit/implicit attitudes. *Linguist. Vanguard* 5:S1. doi: 10.1515/lingvan-2018-0010

Pierrejean, B., and Tanguy, L. (2018). "Towards qualitative word embeddings evaluation: measuring neighbors variation," in *Proceedings of the NAACL-HLT 2018: Student Research Workshop* (New Orleans, LA), 32–39. doi: 10.18653/v1/N18-4005

Preston, D. (2015). "Does language regard vary?," in *Responses to Language Varieties: Variability, Processes, and Outcomes*, eds A. Prikhodkine and D. Preston (Amsterdam: John Benjamins), 1–36.

Purschke, C. (2014). "REACT – Einstellungen als evaluative Routinen in sozialen Praxen," in *Sprechen über Sprache. Perspektiven und neue Methoden der Einstellungsforschung*, eds C. Cuonz and R. Studler (Tübingen: Stauffenburg), 123–142.

Purschke, C. (2015). "REACT – a constructivist theoretic framework for attitudes," in *Responses to Language Varieties: Variability, Processes, and Outcomes*, eds A. Prikhodkine and D. Preston (Amsterdam: John Benjamins), 37–54.

Purschke, C. (2018). "Language regard and cultural practice – variation, evaluation, and change in the German regional languages," in *Language Regard: Methods, Variation, and Change*, eds B. Evans, E. Benson, and J. Stanford (Cambridge: Cambridge University Press), 249–265.

Purschke, C. (2020). "Exploring the linguistic landscape of cities through crowdsourced data," in *Handbook of the Changing World Language Map*, eds S. Brunn and R. Kehrein (Cham: Springer), 1–22. doi: 10.1007/978-3-319-73400-2_220-1

Purschke, C., and Hovy, D. (2019). Lörres, Möppes, and the Swiss. (Re)Discovering regional patterns in anonymous social media data. *J. Linguist. Geogr.* 7, 113–134. doi: 10.1017/jlg.2019.10

Redinger, D. (2010). *Language attitudes and code-switching behaviour in a multilingual educational context: the case of Luxembourg* (Ph.D. thesis). University of York, UK. Available online at: http://etheses.whiterose.ac.uk/1101/

Rodríguez-Penagos, C., Grivolla, J., and Codina Fibá, J. (2012). "A hybrid framework for scalable opinion mining in social media: detecting polarities and attitude targets," in *Proceedings of the 13th Conference of the EACL* (Avignon), 46–52. Available online at: https://www.aclweb.org/anthology/W12-0606

Sieburg, H., and Weimann, B. (2014). "Sprachliche Identifizierungen im luxemburgisch-deutschen Grenzraum," in *Räume und Identitäten in Grenzregionen. Politiken – Medien – Subjekte*, eds C. Wille, R. Reckinger, S. Kmec, and M. Hesse (Bielefeld: transcript), 346–361.

Sirajzade, J., and Schommer, C. (2019). "The LuNa open toolbox for the Luxembourgish language. advances in data mining – applications and theoretical aspects," in *19th Industrial Conference (ICDM 2019): Poster Proceedings* (New York, NY), 1–15. Available online at: http://hdl.handle.net/10993/40407

Somasundaran, S., and Wiebe, J. (2009). "Recognizing stances in online debates," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* (Singapore), 226–234. Available online at: https://www.aclweb.org/anthology/P09-1026

Soukup, B. (2012). Current issues in the social psychological study of 'language attitudes': constructionism, context, and the attitude-behavior link. *Lang. Linguist. Compass* 6, 212–224. doi: 10.1002/lnc3.332

STATEC (2019). *Luxembourg in Figures 2019*. Luxembourg City: STATEC.

Stölben, R. (2019). *Deutsch in Luxemburg. Eine Erhebung zu aktuellen Spracheinstellungen* (Master's thesis). University of Mannheim, Mannheim, Germany.

Strapparava, C., and Mihalcea, R. (2008). "Learning to identify emotions in text," in *Proceedings of the 2008 ACM Symposium on Applied Computing*, 1556–1560. doi: 10.1145/1363686.1364052

Strapparava, C., Valitutti, A., and Stock, O. (2007). Dances with words. *IJCAI* 2007, 1719–1724. doi: 10.5555/1625275.1625554

Taboada, M., and Grieve, J. (2004). "Analyzing appraisal automatically," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 158–161. Available online at: https://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-029.pdf

Tophinke, D., and Ziegler, E. (2014). "Spontane Dialektthematisierung in der Weblogkommunikation: interaktiv-kontextuelle Einbettung, semantische Topoi und sprachliche Konstruktionen," in *Sprechen über Sprache*, eds C. Cuonz and R. Studler (Tübingen: Stauffenburg), 205–242.

Wagner, M. (2012). "Sprachideologien auf Facebook: Diskussionen auf Gruppenseiten über den Sprachgebrauch in Luxemburg," in *Entwicklungen im Web 2.0: Ergebnisse des III. Workshops zur linguistischen Internetforschung*, eds T. Siever and P. Schlobinski (Frankfurt am Main: Peter Lang), 131–150.

Wagner, M. (2013). "Luxembourgish on Facebook: language ideologies and writing strategies", in *Social Media and Minority Languages: Convergence and the Creative Industries*, eds E. H. Gruffydd Jones and E. Uribe-Jongbloed (Bristol: Multilingual Matters), 87–98.

Zenter für d'Lëtzebuerger Sprooch (2019). *D'Lëtzebuerger Orthografie*. Luxembourg City: SCRIPT/ZLS.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/frai.2020.565682

Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation

Ben Gittelson[1]*, Adrian Leemann[2] and Fabian Tomaschek[3]

[1]Internet Institute, Oxford University, Oxford, United Kingdom, [2]Center for the Study of Language and Society, University of Bern, Bern, Switzerland, [3]Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany

This study examines the status of nonmodal phonation (e.g. breathy and creaky voice) in British English using smartphone recordings from over 2,500 speakers. With this novel data collection method, it uncovers effects that have not been reported in past work, such as a relationship between speakers' education and their production of nonmodal phonation. The results also confirm that previous findings on nonmodal phonation, including the greater use of creaky voice by male speakers than female speakers, extend to a much larger and more diverse sample than has been considered previously. This confirmation supports the validity of using crowd-sourced data for phonetic analyses. The acoustic correlates that were examined include fundamental frequency, H1*-H2*, cepstral peak prominence, and harmonic-to-noise ratio.

Keywords: smartphone apps, voice quality, British English, regional variation, phonation

OPEN ACCESS

Edited by:
James N. Stanford,
Dartmouth College, United States

Reviewed by:
Catherine Madill,
The University of Sydney, Australia
Kathryn H Franich,
University of Delaware, United States

***Correspondence:**
Ben Gittelson
ben.gittelson@gmail.com

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: 25 May 2020
Accepted: 07 December 2020
Published: 25 January 2021

Citation:
Gittelson B, Leemann A and Tomaschek F (2021) Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation.
Front. Artif. Intell. 3:565682.
doi: 10.3389/frai.2020.565682

1 INTRODUCTION

Creaky voice—a type of nonmodal phonation resulting from the constriction of the glottis—has inspired a steady stream of frenzied editorials and news pieces in the American and British media over the past decade. *The Spectator* asked whether "creaky voice make[s] you a female yuppie—or an updated Vicki Pollard?" *The Washington Post* claimed that it hurts young women's job prospects, and *AARP The Magazine* warned that it could damage their vocal cords. Despite this attention from the popular media, there has been little scholarly inquiry into the status of nonmodal phonation in British English since the 1980s (Henton and Bladon, 1985). While nonmodal phonation has received more attention in American English, most studies of it have relied on sample sizes of less than 50 participants and have been limited to speakers from specific geographical areas, age groups, and socioeconomic classes. This study attempts to address these gaps by investigating the use of nonmodal phonation in a diverse group of over 2,500 speakers from across the United Kingdom.

1.1 Phonation Types

Phonation types refer to the different methods of producing sound through the vibration of the vocal cords (Keating et al., 2015). These types can be divided into two broad categories: modal and nonmodal. In modal phonation, the vocal folds make full contact during the closed phase of the phonatory cycle; this is not the case in nonmodal phonation (Titze, 1995). Ladefoged (1971) represented phonation types as falling on a one-dimensional articulatory continuum based on the degree of glottal constriction, an assumption that underlies much of the literature on this topic (Yuasa, 2010; Keating et al., 2015; Lancia et al., 2016).

Creaky voice and breathy voice are specific types of nonmodal phonation. In this paper, the umbrella term is used when discussing multiple types of nonmodal phonation simultaneously or

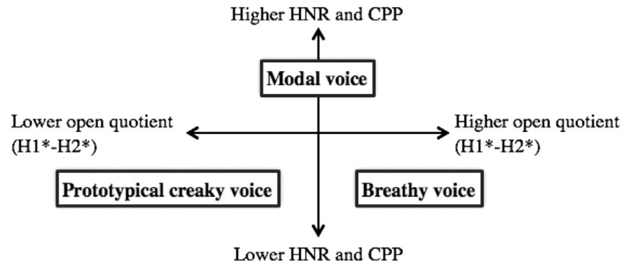

Higher HNR and CPP
Modal voice
Lower open quotient (H1*-H2*)
Higher open quotient (H1*-H2*)
Prototypical creaky voice
Breathy voice
Lower HNR and CPP

FIGURE 1 | Acoustic measures of breathy and creaky nonmodal phonation. Values (higher and lower) are presented relative to those for modal phonation (adapted from Garellek (2012)).

when the acoustic correlates in question would not allow the authors to distinguish between different kinds of nonmodal phonation. When appropriate, the more specific terms are used.

1.2 Acoustic Correlates of Nonmodal Phonation

Multiple acoustic measures may be necessary to fully describe the phonation types on this articulatory continuum, the most common of which are H1-H2 and harmonics-to-noise ratio (HNR). H1-H2 is the difference between the first and second harmonics. The first harmonic is the fundamental frequency, and the second harmonic is the first multiple of the fundamental. H1-H2 serves as a measure of spectral tilt, which is highly correlated with the degree of glottal constriction. In general, lower H1-H2 is associated with creaky voice, while higher H1-H2 occurs with breathy voice (Keating and Esposito, 2006). HNR describes the periodicity of the speech signal; nonmodal phonation results in lower HNR values than modal phonation, as the vibration of the vocal cords is usually less regular (Garellek and Seyfarth, 2016). Cepstral peak prominence (CPP), another measure of periodicity, has also been used to distinguish between modal and nonmodal phonation (Heman-Ackah et al., 2014; Garellek and Seyfarth, 2016). Heman-Ackah et al. (2014) suggested that CPP is a better measure of periodicity than HNR because it does not rely on pitch tracking and is therefore reliable even for very aperiodic signals. The relative values of H1-H2, HNR, and CPP typically associated with modal, breathy, and prototypical creaky voice are represented in **Figure 1** (Garellek, 2012).

Dallaston and Docherty (2020) conducted a systematic review of studies of creaky voice in different varieties of British and American English. They suggested increasing the use of automated acoustic measurement of phonation types, as only one of the nine studies that met their inclusion criteria used this methodology. They argued that using such methods could increase the replicability and scalability of previous conclusions about the status of creaky voice in English, a gap which the present study addresses.

1.3 Sex Differences in the Production of Nonmodal Phonation

Previous work has found differences in the production of nonmodal phonation between men and women using read and spontaneous speech, typically with sample sizes of less than 50 participants and manual coding of phonation types. Henton and Bladon (1985) investigated sex and dialect differences in breathy voice in Received Pronunciation (RP) and Northern British English speakers' open vowels. They selected citation forms of the open vowels /æ/, /ʌ/, /ɒ/, and /ɒ/ from 61 speakers in a preexisting corpus and measured their raw H1-H2 values. The study found that British women produced breathy voice more often than their male counterparts, and that male speakers used creaky voice more frequently than female speakers. Hanson et al. (2001) examined sex differences in the production of open vowels in non-spontaneous speech as part of a larger study on models of phonation types. Specifically, they elicited the vowels /æ/, /ʌ/, and /ɛ/ in carrier phrases from 21 male and 22 female participants. The authors reported two measures of spectral tilt, both corrected for the boosting effects of nearby formants: H1*-A3*, the difference between the amplitude of the first harmonic and the third formant, and H1*-A1, the difference between the amplitude of the first harmonic and the first formant. They concluded that these measures were useful for distinguishing between male and female speakers and that there was wide variation in glottal configuration for both male and female speakers.

Yuasa (2010) investigated sex differences in American English speakers' production of creaky voice. She elicited spontaneous speech from 23 California English speakers, randomly selected 401-word samples from each one, and impressionistically coded occurrences of creaky voice. She found that women produced more creaky voice than men, a finding which was supported by Podesva (2011). However, Dallaston and Docherty (2020)—who included Yuasa (2010) in their systematic review of creaky voice in English—did not find conclusive evidence to substantiate claims of a widespread increase in the use of creak by young American women.

Garellek and Seyfarth (2016) examined acoustic differences between /t/ glottalization and phrasal creak. They used recordings of spontaneous speech from a gender-balanced corpus of 40 adults in Ohio. The researchers identified creaky phonation using preexisting annotations in the corpus and manual inspection. They concluded that linear discriminant models could be used to distinguish between different sources of creaky voice and that CPP was important for identifying this distinction.

1.4 Accent and Ethnicity in the Production of Nonmodal Phonation

The roles of demographic factors such as accent and ethnicity in the production of nonmodal phonation have been studied less extensively than that of sex. However, existing literature suggests that they may be related as well. Within British males, Henton and Bladon (1985) found that RP speakers creaked more than Northern British English speakers. More recently, San Segundo et al. (2019) identified instances of creaky voice in "nearly all" of the 99 Standard Southern British English speakers they studied.

Ethnicity may also play a role in nonmodal phonation. For instance, Alim (2004) linked African American identity to falsetto and "strained" voice qualities. Podesva and Callier (2015) noted that listeners could distinguish African American English speakers from white ones, even in the absence of lexical and syntactic features of African American Vernacular English. They suggested that nonmodal phonation could be responsible for this result.

In this study, we present the first large-scale acoustic analysis of nonmodal phonation for more than 2,500 speakers of British English. We examine how geography, word duration, and social factors such as sex, age, and education level affect the production of nonmodal phonation.

1.5 Hypotheses

We investigated the following hypotheses based on the findings described in **Sections 1.1 through 1.4**. The acoustic correlates used to investigate each hypothesis are described in greater detail in **Section 2.5**.

- Young, highly educated women creak more than men of a similar age (Yuasa, 2010; Podesva, 2011; Melvin and Clopper, 2015).
- These young, highly educated women also creak more than older men and women (Yuasa, 2010; Podesva, 2011; Melvin and Clopper, 2015).
- Men (of all ages) creak more than women (Henton and Bladon, 1985; Foulkes and Docherty, 1999).

2 METHODS

2.1 Recording Method

Phonetic studies typically investigate research questions by having speakers utter words and short sentences in recording chambers at a university. Experiments under laboratory conditions allow researchers maximal control over the context of the recording and the material. However, the recording environment affects the variables of interest (Wagner et al., 2015). For example, these recording chambers do not provide the most naturalistic environment for communication, and this environment typically limits the diversity of the recorded speakers (Henrich et al., 2010; Arnett, 2016). Results are therefore biased toward the group to which scholars have access, which is typically young students. Furthermore, experiments on the university campus limit the number of participants, which ranges from five to 20 in many phonetic studies and as high as 100 or 200 on rare occasions. Small sample sizes lower the probability of detecting a true effect and raise the probability of false positives (Button et al., 2013).

With the rise of the internet, researchers can access a larger and more diverse group of participants than ever before. In addition, speakers can perform experiments in surroundings in which they feel the most comfortable. Though it requires a trade-off with potential variation in recording quality, the use of social media and private recording devices increases researchers' ability to obtain more natural speech from a larger and more diverse group of participants.

In the present paper, we follow this argumentation. In order to record as many speakers as possible from as different backgrounds as possible, we opted to investigate phonation types not in a laboratory but rather by allowing speakers to record their voices on their own phones. To do so, we used the English Dialects App (Leemann et al., 2018), a smartphone program that allows users to record short passages in their native accents and dialects.

2.2 Materials

Before recording the passage, users provided data about their age, gender, education level, mobility, and ethnicity and identified their dialect by placing a pin on the locality that best corresponded to it. They then consented to the privacy agreement shown on the metadata screen. Next, participants were shown the following recording instructions: "Please record your voice in a quiet place. Hold your device approximately 6 inches/15 cm from your mouth. Please use your regional accent or dialect and speak in the way you would talk to your friends from home." After reading these instructions, users created and uploaded recordings, in which they read a passage from "The Boy Who Cried Wolf" sentence by sentence (Deterding, 2006). The user interface then prompted speakers to self-declare their dialect by placing a pin on a map and to provide other metadata, such as age and gender. After recording, users were able to click "play" to hear their recordings and were able to re-record them if they were unsatisfied. Once satisfied, they could then navigate an interactive map where their and others' recordings were uploaded. Upon submitting the recordings, users were shown the following notification: "by clicking 'start recording', you agree with our privacy policy, see info tab." None of the information elicited–about accent, age, gender, et cetera–allows for identification of a user in the database, either individually or when considered in combination. Please see Leemann et al. (2018) for more detail about the corpus structure and demographic makeup of the speakers.

2.3 Speakers

Because the original data did not contain any participant identifiers, we created a participant ID using latitude, longitude, age, gender, education level and ethnicity. This yielded 2,931 participants. On that basis, we found that some participants had recorded the same stimuli more than once in different sessions. We excluded those participants (N = 159) from the analysis, leaving 2,772 speakers. We also excluded speakers from the analysis who had not yet finished school (N = 208), leaving us with a total of N = 2,564 on whom acoustic analyses were performed.

2.4 Signal Processing

The words used in this study were selected from the 10 sentences in "The Boy Who Cried Wolf." Words were considered if and only if they consisted solely of vowels and voiced consonants, i.e. sonorants or phonemically voiced obstruents. The sole exception was the /h/ in "however."

To narrow down this word list, recordings were automatically segmented using WebMAUS (Kisler et al., 2017), which aligned recordings with the corresponding sentence's orthographic and phonemic transcription. The SAMPA phonemic transcriptions of these utterances were generated using the MAUS grapheme-to-phoneme (G2P) model and were manually verified before use. For each word, a random subset of 25 recordings was examined by hand to ensure that the forced alignments were correct. Words were only selected for analysis if at least 24 out of the 25 recordings were correctly aligned. This process led to a final list of six words in utterance-initial, medial, and final positions: "being," "boy," "gave," "however," "one," and "while." These words were then extracted from their respective utterances using a Praat (Boersma and Weenink, 2020) script and the TextGrids generated by WebMAUS.

We applied several methods to ensure that the extracted recordings actually contained the words of interest. We flagged words automatically on the basis of duration comparison and a calculation of zero crossing. The accuracy of the word boundaries in these recordings was then manually verified. We furthermore trimmed white spaces in an automatic procedure using the amplitude envelope as a measure of signal onset and offset. After this procedure, we analyzed all six words for 1,958 speakers, five words for 423 speakers, four words for 103 speakers, three words for 24 speakers, two words for 24 speakers, and one word for 32 speakers.

2.5 Data Analysis

As noted in **Section 1.2**, a wide variety of acoustic correlates have been used to study nonmodal phonation in previous literature. All commonly used metrics were investigated in this study to ensure comparability with prior work. These included HNR35, H1*-H2*, CPP, and F0. We used the corrected H1*-H2* rather than the raw H1-H2 measure to account for the fact that formants raise the amplitudes of nearby harmonics, making it difficult to compare H1-H2 values across different vowels (Hanson et al., 2001). This study used the correction formula described by Iseli et al. (2007) and implemented in VoiceSauce, which subtracts the amount by which the formants raise the harmonics to recover the magnitudes of the source spectrum. HNR35 is the harmonic-to-noise ratio measured between zero and 3,500 Hz. Each of the measures was calculated for 10 time steps across the word, and the mean value of those measurements was used for analysis. Numerical predictors were z-scaled to allow for comparability of the effect sizes. The following variables were used as predictors in our analyses:

- Gender (reference "female").
- Age (mean = 34.3, sd = 14.8).
- Latitude and longitude of the location where the recording was performed. Pilot analyses revealed no effect of latitude and longitude, so these variables were omitted in the final models.
- Education level. Speakers were asked to select the degree of their education level. Possible answers were, in decreasing rank: "Higher Education (BA, BSc, MA etc., PGCE) and professional/vocational equivalents"; "A levels, Bac, vocational level 3 etc."; "5 GCSE grade A*-C, 5 O-Levels, vocational level 2 etc."; "Fewer than 5 GCSE grade A*-C, or fewer than 5 O-Levels," "unknown," and "No qualifications." We transformed education level into a ranked scale, where higher values corresponded to higher education levels and vice versa. It is possible that education level is strongly correlated with age, posing a problem of collinearity in the model. Although the Spearman's rank-correlation between education level and age was significant in the present study, it was not strong enough to be harmful to the regression analysis ($\rho = 0.23$, $p < 0.001$).
- The duration of the word. Word duration was log-transformed to reduce overly strong influence from outliers.
- Mean fundamental frequency in the extracted word (F0).

TABLE 1 | Linear mixed-effects regression summary table for F0. Absolute t-values larger than 2 are regarded to indicate significance and are highlighted in bold.

	Estimate	Std. Error	t-value
(Intercept)	221.267	7.527	**29.398**
Word duration	−2.468	0.350	**−7.073**
Gender = Male	−85.131	0.940	**−90.742**
Speaker age	−7.855	0.660	**−11.894**
Education level	−2.954	0.411	**−7.172**
Gender = Male : Speaker age	5.929	0.935	**6.341**

We used linear mixed-effects regression (LMER, Bates et al. (2015)) to investigate the relationship between these predictors and our measures of nonmodal phonation. We accounted for systematic effects of speakers by including random intercepts for subjects and for systematic effects of items by including random intercepts for words. Given that random intercepts shrink strong outliers more towards the mean than those already close to the mean, an estimate of *p*-values is not possible. Rather, the significance of LMER models is derived from the t-value. Absolute t-values (with t = estimate/standard error) larger than 2 are regarded to indicate a significant effect. We also included random slopes by participant. The predictors that were included as random slopes are indicated below in the Results section.

We performed an exploratory top-down and bottom-up statistical analysis, comparing different models using AIC and inspecting the significance of predictors and interactions. The final model structure included a main effect for word duration, gender, education level and an interaction between Gender and Speaker Age. In addition, F0 was used as a main effect in models fitting HNR35, H1*-H2*, and CPP.

3 RESULTS

3.1 The F0 Measure

We tested models with three different F0 trackers: Snack (Sjölander, et al., 1998), STRAIGHT (Kawahara et al., 1998), and SHR (Sun, 2002). We used the output from each of these trackers as dependent variables and found that STRAIGHT

TABLE 2 | Linear mixed-effects regression summary table for HNR35. Absolute t-values larger than 2 are regarded to indicate significance and are highlighted in bold.

	Estimate	Std. Error	t-value
(Intercept)	36.871	1.374	**26.823**
F0	2.661	0.096	**27.787**
Word duration	0.678	0.058	**11.777**
Gender = Male	−3.750	0.239	**−15.705**
Speaker age	0.608	0.136	**4.469**
Education level	−0.244	0.087	**−2.796**
Gender = Male : Speaker age	−0.808	0.199	**−4.068**

yielded the best model fit (total AIC decrease of 5,280 between Snack and Straight, with SHR in between). We included word duration as correlated random slopes by participant.

The final model, summarized in **Table 1**, found lower F0 values in longer words and for male, older, and less educated speakers. The first row of **Figure 2** visualizes the results, where F0 is represented on the *y*-axis and the predictor on the *x*-axis. Furthermore, the significant gender and age interaction indicates that the effect of age was smaller in male speakers than in female speakers.

3.2 The HNR35 Measure

Testing models with HNR05, HNR15, HNR25 and HNR35 as dependent variables against each other, we found that the HNR35 measure resulted in the best model fit (AIC decrease of 7,031 between HNR05 and HNR35, with HNR15 and HNR25 in between). In addition to the predictors presented above, we included F0 as a predictor for HNR. In spite of its significant correlation with all other predictors, no effects of suppression, i.e. changes in signs, and enhancement, i.e. anti-conservative p-values, were present in this and all of the following models, which is why we regard its inclusion as safe (cf. Tomaschek et al. (2018)). We included F0 as correlated random slopes by participant. The model failed to converge with word duration as random slopes.

The second row of **Figure 2** illustrates the results. We found significantly lower HNR values in words with lower mean F0, in shorter words, in male speakers relative to female speakers, and in speakers with higher education levels. The significant gender and age interaction indicates that female speakers' HNR values increase as they get older, while this effect is reversed for male speakers. Note that the effect size is strongest for F0 and smaller by an order of 10 for all other predictors. This difference in effect size is mirrored in the other measures.

3.3 The H1*-H2* Measure

We included F0 as uncorrelated random slopes by participant to the model fitting H1*-H2*. The model failed to converge with word duration as random slopes. The third row of **Figure 2** shows the results for H1*-H2*. We found lower H1*-H2* values in words with lower mean F0, in male speakers than in female speakers, and in longer words than in shorter words. The significant gender-age interaction indicates that older male speakers have lower H1*-H2* values. No effect of age was found for female speakers. Also, no effect was found for education level. Overall, the size of the effects is comparatively smaller for H1*-H2* than for F0 and HNR.

3.4 The CPP Measure

We included F0 as correlated random slopes by participant to the model fitting CPP. The model failed to converge with word duration as random slopes. The bottom row of **Figure 2** displays the effects of CPP. Pitting the CPP measure against our predictors, we found significantly lower CPP values associated with higher F0 values and with older age (see **Table 4**). None of the other effects yielded significance.

4 DISCUSSION

This discussion will consider the effect of demographic variables (sex, age, and education level) and F0 on the production of nonmodal phonation in British English. The results indicate that male speakers, older speakers, and more educated speakers produce more nonmodal phonation than female, younger, and less educated speakers and that more nonmodal phonation is associated with lower F0. We will end the discussion with a note on limitations.

4.1 Sex

Overall, our findings demonstrated that male speakers produced more creaky voice than female speakers. This was borne out in the fact that men had lower HNR than women, where lower HNR is associated with less periodicity in the speech signal and more nonmodal phonation. H1*-H2*, which measures the difference between the first and second harmonics, was also lower for men than for women, confirming that men creaked more than women. These findings are consistent with Henton and Bladon (1985) and Foulkes and Docherty (1999), who, for a subset of UK speakers, found that male speakers tended to produce more creaky voice than female speakers. In American English, two relatively recent studies (Yuasa, 2010; Podesva, 2011) demonstrated that women creaked more than men; the present results indicate that this phenomenon is not present in British English.

4.2 Age

Older speakers generally produced more creaky voice than younger speakers, though this effect was modulated by sex. Overall, older participants exhibited lower HNR35 and H1*-H2* values than younger ones. These findings on HNR are consistent with research on presbyphonia, or age-related changes to the vocal tract. For instance, Lortie et al. (2015) similarly found that older speakers tended to have lower HNR values than younger and middle aged ones.

Further investigation revealed that this relationship differed between sexes. For men, HNR35 and H1*-H2* followed the overall trend of decreasing with age, indicating that older men produced more creaky voice than younger ones. However, the opposite was true for women. This finding contrasts with that of Ferrand (2002), who found that elderly females had substantially lower HNR35 values than the two other age cohorts they compared to.

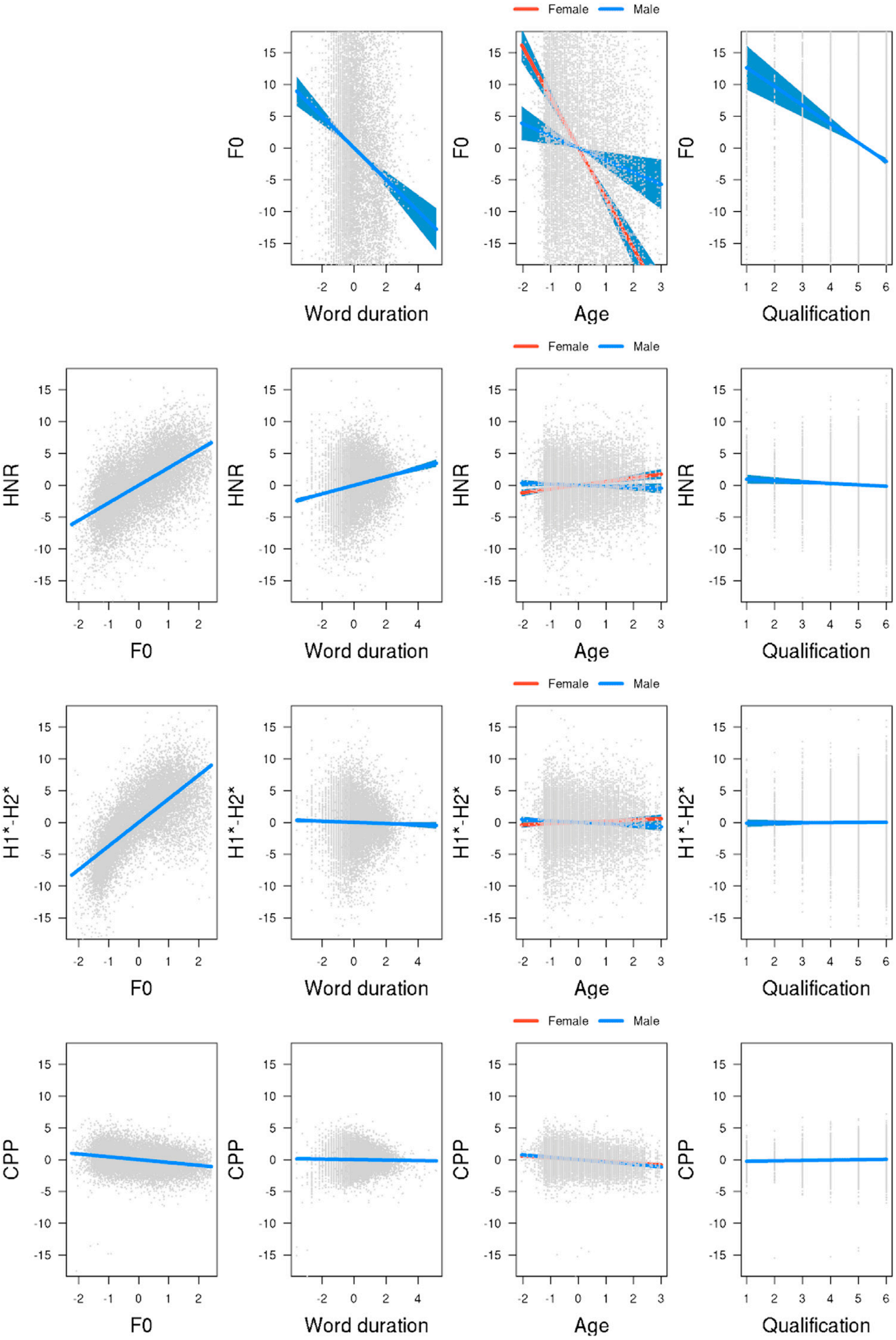


FIGURE 2 | Model estimates for the four measures of modal voice. First row: F0, second row: HNR, third row: H1*-H2*, bottom row: CPP. Predictors are illustrated in the columns (adapted from Garellek (2012)).

4.3 Education Level

The results for education level suggest that more educated speakers produced more creaky voice. Specifically, they exhibited both lower F0 and HNR values. Lower HNR indicates increased likelihood of nonmodal phonation—either creaky or breathy—while lower F0 suggests that the speakers produce creaky voice. These findings mirror those found for highly educated women in the U.S. (as described in **Section 1.3**).

TABLE 3 | Linear mixed-effects regression summary table for H1*-H2*. Absolute t-values larger than 2 are regarded to indicate significance and are highlighted in bold.

	Estimate	Std. Error	t-value
(Intercept)	2.695	0.368	**7.320**
F0	4.563	0.113	**40.134**
Word duration	−0.115	0.048	**−2.415**
Gender = Male	−0.740	0.173	**−4.274**
Speaker age	0.186	0.095	1.962
Education level	−0.032	0.063	−0.507
Gender = Male : Speaker age	−0.303	0.141	**−2.140**

Voice disorders, such as dysphonia, may also help explain this association between education level and the production of creaky voice. Roy et al. (2005) reported that the lifetime prevalence of self-reported voice disorders could be as high as 29.9 percent in the general population, while Bhattacharyya (2014) found that it was closer to 7.6 percent. Occupational voice users, such as teachers and singers, report a high prevalence of such disorders and may tend to be more highly educated (Timmermans et al., 2002). Timmermans et al. (2002) reported a statistically significant difference in acoustic measures of voice disorders between a group of occupational voice users and a control group. These measures included jitter and the highest possible F0 produced by each subject.

However, other investigations such as Niebudek-Bogusz et al. (2006) and Lehto et al. (2006) have not established a significant relationship between self reports of voice disorders in occupational voice users and objective acoustic measures of these disorders. Furthermore, the effect found in this investigation was a significant relationship between education and HNR35; this relationship was not significant for CPP, which Heman-Ackah et al. (2014) indicated was a better acoustic measure of dysphonia.

4.4 F0

Findings on sex and age differences in F0 align with previous research in this area, suggesting that the large-scale automated F0 tracking produced valid results. For instance, men exhibited a lower F0 than women. We also found that women's F0 decreased with age, an effect that is consistent with research on presbyphonia (Linville and Fisher, 1985; Bruzzi et al., 2017). Lower F0 was generally associated with more nonmodal phonation, even when sex was taken into account. Lower values of H1*-H2*, HNR, and CPP, all of which indicate an increased likelihood of nonmodal phonation, were associated with lower F0. This may occur because lower frequencies of vocal fold vibration make it more likely that phonation becomes irregular, and thus creaky (Keating et al., 2015).

4.5 Comparison of Measures

We found that the strongest effects of gender, age and education level could be observed for F0, followed by HNR35, H1*-H2*. Our predictors showed the weakest effects for CPP. The small effect sizes for the non-F0 measures could be a result of the fact that F0 was used as a predictor in these models, accounting for a large proportion of the variance.

4.6 Limitations

Mobile phone recordings allowed for the development of a large and diverse data set, but this data collection method is not without its limitations. For example, European privacy regulations prohibited the collection of information about the sampling rate, bit rate, and type of encoding used by the different smartphone devices. Unknown recording conditions may have have also negatively impacted signal quality, as signals with more noise produce less reliable acoustic analyses and forced alignments. Despite a lack of a control of signal type, we still found the same patterns of phonation type variation across the United Kingdom as in previous studies that used controlled acoustic measurements. Crowd-sourced data requires a trade-off between a relative lack of control of signal quality and large, diverse data sets.

A number of studies have demonstrated that smartphone devices produce similar acoustic measurements to those found in laboratory recordings. Smartphone recordings have been shown to be sufficient for formant analysis (Decker and Nycz, 2011). Grillo et al. (2016) demonstrated that various Apple and Samsung smartphones produced similar F0, HNR, and CPP measurements to laboratory-quality microphones. A more recent study by Jannetts et al. (2019) considered four different devices (Samsung Galaxy S8+, iPhone 6s, iPhone 7, and Samsung Galaxy J3) and their effects on acoustic parameters. When compared to a reference microphone (Neumann U89i), they reported that acoustic parameters could be measured with smartphones with varying degrees of reliability. F0 and CPP, for example, provided relatively robust measures, while jitter and shimmer, which were not included in this study, did not. Jannetts et al. (2019) found that Samsung phones produced F0 values that were slightly higher than the reference measurements, while the Apple phones were slightly too low, though these errors never exceeded 2Hz. For CPP measures, all phones revealed somewhat lower values than the reference measures (Samsung c. -0.5dB; Apple -08 to -1dB). Note, though, that the authors state that these errors are so low that "their practical relevance is probably limited." Furthermore, CPP measures did not provide significant effects in our statistical models. Unfortunately, Jannetts et al. (2019) did not study the devices' effects on HNR parameters.

TABLE 4 | Linear mixed-effects regression summary table for CPP. Absolute t-values larger than 2 are regarded to indicate significance and are highlighted in bold.

	Estimate	Std. Error	t-value
(Intercept)	19.310	0.276	**69.924**
F0	−0.468	0.040	**−11.678**
Word duration	−0.039	0.026	−1.497
Gender = Male	0.082	0.096	0.854
Speaker age	−0.277	0.052	**−5.369**
Education level	0.064	0.033	1.907
Gender = Male : Speaker age	−0.081	0.076	−1.062

As an anonymous reviewer has pointed out, voiced plosives and glides create F0 contours (Ladd and Schmid, 2018), which will influence the HNR values. These kind of dynamic changes are inherent to the natural speech that was the focus of the current study. As a consequence, it is almost impossible to extract phonetic signals with constant F0. We therefore rather focus on a large number of samples with dynamic F0, such that any effects of dynamic transitions will be averaged across words and speakers in a large data set like the present one. Our results mirror the findings from studies that used highly controlled recording environments and measurements from vowels, which suggests that this was a valid approach.

We did not collect data on the socioeconomic or health status of our subjects due to privacy concerns, and these variables could have impacted our findings, particularly to the extent that they may be related to dysphonia. For example, Cohen et al. (2012) found that a plurality of dysphonia-related health insurance claims in the United States were filed by workers in lower paid manufacturing jobs. Dysphonia also frequently co-occurs with other health conditions, such as bronchitis and pneumonia (Cohen et al., 2012). Future studies should consider whether and how to collect such data at scale and its relationship with the production of nonmodal phonation.

5 CONCLUSION

Further research should attempt to address these concerns and consider the perceptual and phonological implications of this study's findings. A natural progression of this work would be to conduct a perceptual study of phonation type measures. That is, do listeners perceive a difference in phonation type if words or utterances are resynthesized with different values for F0, H1*-H2*, CPP, HNR, etc.? Future studies should also consider the effect of phrase position on nonmodal phonation, as it has been suggested that creaky voice often occurs phrase-finally (Henton, 1986; Podesva and Callier, 2015).

The results of this study indicate that conclusions about the interaction of age, sex and nonmodal phonation from the 1980s and 1990s with small and geographically limited samples hold true for a large and demographically diverse group of current-day British English speakers. The use of crowd-sourced big data also allowed this study to uncover previously unobserved effects, such as a relationship between nonmodal phonation and education level. Taken as a whole, these results support the validity of using big data in phonetic studies and demonstrate that other researchers should use such data sets to confirm or challenge previous conclusions about the acoustic properties of British English speech.

DATA AVAILABILITY STATEMENT

The data and analysis supporting the conclusions of this article can be found at https://osf.io/bvyt2/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Zurich cantonal ethics committee (http://www.kek.zh.ch/internet/gesundheitsdirektion/kek/de/home.html) and followed the accompanying federal laws on experimentation on humans in Switzerland, where the app was developed (http://www.admin.ch/opc/de/classified-compilation/20061313/index.html). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BG: conceptualization, literature review, methodology, data cleaning, data analysis, writing, editing. AL: conceptualization, data curation, investigation, methodology, project administration, resources, supervision, visualization, writing, editing. FT: conceptualization, methodology, statistical analyses, visualization, writing, editing.

FUNDING

This research was funded in part by the Deutsche Forschungsgemeinschaft (Research Unit FOR2373 'Spoken Morphology', Project 'Articulation of morphologically complex words', BA 3080/3-1). Deutsche Forschungsgemeinschaft (Research Unit FOR2373 Spoken Morphology, Project Articulation of morphologically complex words), BA 3080/3-2.

ACKNOWLEDGMENTS

We thank Yang Li and Nianheng Wu, who provided insight and expertise that greatly assisted this research, although they may not agree with all of the interpretations and conclusions of this paper. We further wish to thank David Britain (Bern), Tam Blaxter (Cambridge), Marie-José Kolly (Republik), and Daniel Wanitsch (ibros.ch) for co-developing the English Dialects App. The app provided the dataset the current paper is based on.

REFERENCES

Alim, H. S. (2004). *You know my steez: an ethnographic and sociolinguistic study of styleshifting in a Black American speech community.* Durham, N.C.: Duke University Press for the American Dialect Society.

Arnett, J. J. (2016). The neglected 95%: why american psychology needs to become less american. *Am Pyscho*. 63, 602–614. doi:10.1037/0003-066X.63.7.602

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statistical Software*. 67, 1–48. doi:10.18637/jss.v067.i01

Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the United States. *Laryngoscope*. 124, 2359–2362. doi:10.1002/lary.24740

Boersma, P., and Weenink, D. (2020). Praat: doing phonetics by computer. [Computer program]. Version 6.0.37. Available at: http://www.praat.org/ (Accessed January 3, 2020).

Bruzzi, C., Salsi, D., Minghetti, D., Negri, M., Casolino, D., and Sessa, M. (2017). Presbiphonya. *Acta biomed*. 88, 6–10. doi:10.23750/abm.v88i1.5266

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci*. 14, 365–376. doi:10.1038/nrn3475

Cohen, S. M., Kim, J., Roy, N., Asche, C., and Courey, M. (2012). Prevalence and causes of dysphonia in a large treatment-seeking population: prevalence and Causes of Dysphonia. *The Laryngoscope*. 122, 343–348. doi:10.1002/lary.22426

Dallaston, K., and Docherty, G. (2020). The quantitative prevalence of creaky voice (vocal fry) in varieties of English: a systematic review of the literature. *PLos One*. 15, e0229960. doi:10.1371/journal.pone.0229960

Decker, P. D., and Nycz, J. (2011). For the record: which digital media can be used for sociophonetic analysis? University of Pennsylvania Working Papers in Linguistics. 17, 11. doi:10.6084/M9.FIGSHARE.1230096.V1

Deterding, D. (2006). The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *J. Int. Phonetic Assoc*. 36, 187. doi:10.1017/S0025100306002544

Ferrand, C. T. (2002). Harmonics-to-Noise ratio. *J. Voice*. 16, 480–487. doi:10.1016/S0892-1997(02)00123-6

Foulkes, P., and Docherty, G. J. (1999). *Urban voices: accent studies in the British Isles*. London: New York: Oxford University Press.

Garellek, M., and Seyfarth, S. (2016). *Acoustic differences between English /t/ glottalization and phrasal creak*. Proc. Interspeech 2016, 1054–1058. doi:10.21437/Interspeech.2016-1472

Garellek, M. (2012). The timing and sequencing of coarticulated non-modal phonation in English and white Hmong. *J. Phonetics*. 40, 152–161.

Grillo, E. U., Brosious, J. N., Sorrell, S. L., and Anand, S. (2016). Influence of smartphones and software on acoustic voice measures. *Int. J. Telerehabilitation*. 8, 9–14. doi:10.5195/IJT.2016.6202

Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y., and Slifka, J. (2001). Towards models of phonation. *J. Phonetics*. 29, 451–480. doi:10.1006/jpho.2001.0146

Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., et al. (2014). Quantifying the cepstral peak prominence, a measure of dysphonia. *J. Voice*. 28, 783–788. doi:10.1016/j.jvoice.2014.05.005

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not weird. *Nature*. 466, 29. doi:10.1038/466029a

Henton, C., and Bladon, R. (1985). Breathiness in normal female speech: inefficiency versus desirability. *Lang. Commun*. 5, 221–227. doi:10.1016/0271-5309(85)90012-6

Henton, C. G. (1986). Creak as a sociophonetic marker. *J. Acoust. Soc. Am*. 80, S50. doi:10.1121/1.2023837

Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am*. 121, 2283–2295. doi:10.1121/1.2697522

Jannetts, S., Schaeffler, F., Beck, J., and Cowen, S. (2019). Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types. *Int. J. Lang. Commun. Disord*. 54, 292–305. doi:10.1111/1460-6984.12457

Kawahara, H., de Cheveigné, A., and Patterson, R. (1998). An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite, The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, November 30–December 4, 1998.

Keating, P. A., and Esposito, C. (2006). Linguistic Voice Quality. *UCLA Working Papers in Phonetics* 105, 85–91.

Keating, P., Garellek, M., and Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice, 18th International Congress of Phonetic Sciences At: Glasgow, Scotland, August 10–14, 2015.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Comput. Speech Lang*. 45, 326–347. doi:10.1016/j.csl.2017.01.005

Ladd, D. R., and Schmid, S. (2018). Obstruent voicing effects on f0, but without voicing: phonetic correlates of Swiss German lenis, fortis, and aspirated stops. *J. Phonetics*. 71, 229–248. doi:10.1016/j.wocn.2018.09.003

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

Lancia, L., Voigt, D., and Krasovitskiy, G. (2016). Characterization of laryngealization as irregular vocal fold vibration and interaction with prosodic prominence. *J. Phonetics*. 54, 80–97. doi:10.1016/j.wocn.2015.08.001

Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English Dialects App: the creation of a crowdsourced dialect corpus. *Ampersand*. 5, 1–17. doi:10.1016/j.amper.2017.11.001

Lehto, L., Laaksonen, L., Vilkman, E., and Alku, P. (2006). Occupational voice complaints and objective acoustic measurements—do they correlate? *Logopedics Phoniatrics Vocology*. 31, 147–152. doi:10.1080/14015430600654654

Linville, S. E., and Fisher, H. B. (1985). Acoustic characteristics of women's voices with advancing age. *J. Gerontol*. 40, 324–330. doi:10.1093/geronj/40.3.324

Lortie, C. L., Thibeault, M., Guitton, M. J., and Tremblay, P. (2015). Effects of age on the amplitude, frequency and perceived quality of voice. *Age (Dordr)*. 37, 117. doi:10.1007/s11357-015-9854-1

Melvin, S., and Clopper, C. G. (2015). Gender variation IN creaky voice and fundamental frequency. Master's Theses. Ohio: The Ohio State University.

Niebudek-Bogusz, E., Fiszer, M., Kotylo, P., and Sliwinska-Kowalska, M. (2006). Diagnostic value of voice acoustic analysis in assessment of occupational voice pathologies in teachers. *Logopedics Phoniatrics Vocology*. 31, 100–106. doi:10.1080/14015430500295756

Podesva, R. J., and Callier, P. (2015). Voice quality and identity. *Annu. Rev. Appl. Ling*. 35, 173–194. doi:10.1017/S0267190514000270

Podesva, R. J. (2011). Gender and the social meaning of non-modal phonation types. *Annual Meeting of the Berkeley Linguistics Society*. 37, 427. doi:10.3765/bls.v37i1.832

Roy, N., Merrill, R. M., Gray, S. D., and Smith, E. M. (2005). Voice disorders in the general population: prevalence, risk factors, and occupational impact:. *The Laryngoscope*. 115, 1988–1995. doi:10.1097/01.mlg.0000179174.32345.41

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., and Kavanagh, C. (2019). The use of the Vocal Profile Analysis for speaker characterization: methodological proposals. *J. Int. Phonetic Assoc*. 49, 353–380. doi:10.1017/S0025100318000130

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Sjölander, R., and Granström, B. (1998). "Web-based educational tools for speech technology," in Proc of ICSLP98, 5th Intl Conference on Spoken Language Processing. Sydney, Australia, 3217–3220.

Sun, X. (2002). "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, May 13–17, 2002 (IEEE).

Timmermans, B., De Bodt, M., Wuyts, F., Boudewijns, A., Clement, G., Peeters, A., et al. (2002). Poor voice quality in future elite vocal performers and professional voice users. *J. Voice*. 16, 372–382. doi:10.1016/S0892-1997(02)00108-X

Titze, I. (1995). "Definitions and nomenclature related to voice quality,". in *Vocal fold physiology: voice quality control*. O. Fujimura and M. Hirano (San Diego, CA: Singular Publishing Group), 335–342.

Tomaschek, F., Hendrix, P., and Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *J. Phonetics*. 71, 249–267. doi:10.1016/j.wocn.2018.09.004

Wagner, P., Trouvain, J., and Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *J. Phonetics*. 48, 1–12. doi:10.1016/j.wocn.2014.11.001

Yuasa, I. P. (2010). Creaky voice: a new feminine voice quality for young urban-oriented upwardly mobile AMERICAN women? *Am. Speech*. 85, 315–337. doi:10.1215/00031283-2010-018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

CORRECTION
published: 19 May 2021
doi: 10.3389/frai.2021.692064

Corrigendum: Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation

Ben Gittelson[1]*, *Adrian Leemann*[2] *and Fabian Tomaschek*[3]

[1] Internet Institute, Oxford University, Oxford, United Kingdom, [2] Center for the Study of Language and Society, University of Bern, Bern, Switzerland, [3] Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany

Keywords: smartphone apps, voice quality, British English, regional variation, phonation

OPEN ACCESS

Approved by:
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

****Correspondence:***
Ben Gittelson
ben.gittelson@gmail.com

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *07 April 2021*
Accepted: *08 April 2021*
Published: *19 May 2021*

Citation:
Gittelson B, Leemann A and Tomaschek F (2021) Corrigendum: Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation. Front. Artif. Intell. 4:692064. doi: 10.3389/frai.2021.692064

A Corrigendum on

Using Crowd-Sourced Speech Data to Study Socially Constrained Variation in Nonmodal Phonation
by Gittelson, B., Leemann, A., and Tomaschek, F. (2021). Front. Artif. Intell. 3:565682. doi: 10.3389/frai.2020.565682

In the original article, we neglected to include the funder "Deutsche Forschungsgemeinschaft (Research Unit FOR2373 Spoken Morphology, Project Articulation of morphologically complex words), BA 3080/3-2."

In addition, a sentence was omitted from the **Acknowledgments** section. The section now reads:

ACKNOWLEDGMENTS

We thank Yang Li and Nianheng Wu, who provided insight and expertise that greatly assisted this research, although they may not agree with all of the interpretations and conclusions of this paper. We further wish to thank David Britain (Bern), Tam Blaxter (Cambridge), Marie-José Kolly (Republik), and Daniel Wanitsch (ibros.ch) for co-developing the English Dialects App. The app provided the dataset the current paper is based on.

Finally, we did not provide a link to the supporting data in the original **Data Availability Statement**. A correction has been made to the section, as seen below:

DATA AVAILABILITY STATEMENT

The data and analysis supporting the conclusions of this article can be found at https://osf.io/bvyt2/.

The authors apologize for these errors and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/frai.2020.539920

Digital Articulation: Examining Text-Based Linguistic Performances in Mobile Communication Through Keystroke-Logging Analysis

*Joel Schneier**

Department of Writing and Rhetoric, University of Central Florida, Orlando, FL, United States

This study examines how text-based mobile communication practices are performatively constructed as individuals compose messages key-by-key on virtual keyboards, and how these *synchronous performances* (Mobile interface theory: embodied space and locative media. New York, NY: Routledge) reflect the iterative process of constructing and maintaining interpersonal relationships. In doing so, this study reports on keystroke-logging analysis (see Writ. Commun. 30, 358–392) in order to observe how participants (*N* = 10) composed text as part of everyday mobile communication for the period of one week, subsequently producing 179,996 individual keystroke log-file records. Participants used LogKey, a virtual keyboard application made exclusively for this study to run on the Android mobile operating system. Analysis of keystroke log-file data suggest that timing processes of composing text-messages may differ as participants messaged with different categories of interlocutors, composed on different communication applications, and composed paralinguistic features—such as variants of *Lol* and *Haha* Thurlow and Brown, (Discourse Anal. Online, 2003, 1, 1); Tagg, (Discourse of text messaging. 2012, Bloomsbury, UK)—at different turn-taking positions. This evidence suggests that keystroke-logging methods may contribute to understanding of how individuals manage interpersonal relationships in real-time (Please reply! the replying norm in adolescent SMS communication," in The inside text: social, cultural and design perspectives on SMS. (Norwell, MA: Springer), 53–73); (Beyond genre: closings and relational work in texting," in Digital discourse: language in the new media. (Oxford: Oxford University Press), 67–85), and suggests future direction for methodologically studying linguistic performances as part of text-based mobile communication.

Keywords: keystroke analysis, mobile communication, paralinguistic cues, digital articulation, text messaging, computational sociolinguistics

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Heike Wiese,
Humboldt University of Berlin,
Germany
Ryan L. Boyd,
Lancaster University, United Kingdom
Barbara Plank,
IT University of Copenhagen, Denmark

****Correspondence:***
Joel Schneier
joel.schneier@ucf.edu

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence.

Received: *03 March 2020*
Accepted: *01 December 2020*
Published: *25 January 2021*

Citation:
Schneier J (2021) Digital Articulation: Examining Text-Based Linguistic Performances in Mobile Communication Through Keystroke-Logging Analysis. Front. Artif. Intell. 3:539920. doi: 10.3389/frai.2020.539920

INTRODUCTION

The increasing ubiquity of mobile technology in recent decades has given rise to forms of sociolinguistic research that have explored how text-based language may be may be used as part of everyday interpersonal discourses in text-messaging (see Ling and Yttri, 2001; Thurlow and Brown, 2003; Spilioti, 2011; Tagg, 2012), as well as perform social identities in social media (see Pavalanathan and Eisenstein, 2015a; Jones, 2015). Computational methods, such as Grieve, Nini, and

Guo (2017) and Pavalanathan and Eisenstein (2015b), have similarly demonstrated the value in examining how text-based linguistic features may be transmitted and diffused across online social networks to be made part of individual and social performances (Coupland, 2007; Androutsopoulos, 2014b) via broadcast mechanisms available in everyday mobile telephony, particularly social networking sites such as Twitter. In this way, tracing how language may be traded in through online interactions, conducted at the touch of a screen, has provided sociolinguists a front-row seat to witness language diffusion as it occurs in real-time across the Twitter-verse (see Jones, 2015; Grieve et al., 2017).

While these methods have yielded insights into how linguistic forms may be diffused across online social networks, as well as suggest how individuals may adopt and use newly enregistered features for performing social identities, lost among these methods is the notion of a flesh-and-blood performer of language. After all, the underlying assumption in quantitatively and computationally examining the ‘firehose’ of linguistic data on Twitter is that individual humans were responsible for composing said tweets, and that such performances reflect real-life social and linguistic meaning (Brock, 2020; Eisenstein, 2013; Jones, 2015). This study therefore asks, *how can we examine how individuals use mobile technology to compose text-based linguistic features in real-time?* And, *how do the timing processes of composing these linguistic features through mobile media demonstrate how individuals perform social identities?* I therefore present a methodology that examines how text-based mobile communication practices are composed in real-time through keystroke analysis, and suggest that such a methodology, alongside established computational methods to examine the large scores of public text-based data, contributes to a stronger understanding of how mobily-mediated linguistic performances meaningfully unfold in individuals’ everyday lives.

This study therefore reports on keystroke-logging data (see Leijten and Van Waes, 2013) as part of observations of how participants ($N = 10$) composed text as part of everyday mobile communication for the period of one week, subsequently producing 179,996 individual keystroke log-file records. Participants used LogKey, a virtual keyboard application developed exclusively for this study to run on the Android mobile operating system. This small study therefore served as a preliminary test of the feasibility of using LogKey to conduct keystroke analysis on individuals’ own mobile devices, as well examine this study’s stated questions. Analysis of keystroke log-file data obtained from this study yielded insights suggesting that timing processes of composing text may differ as participants messaged with different interlocutors, compose text for different mobile applications, and composed paralinguistic cues—such as variants of *Lol* and *Haha* (Thurlow and Brown, 2003; Tagg, 2012)—at different discursive positions in a text-message. I argue that the findings from this small study, while not generalizable, may contribute to stronger understandings of how individuals manage interpersonal relationships in real-time through composing, sending, and receiving text through mobile communication (Ling and Yttri, 2001; Thurlow and Brown, 2003; Laursen, 2005; Spilioti, 2011). Further, this study’s use of LogKey to examine individuals’ everyday text-based linguistic performances in mobile communication over the period of one week represents one of the first—if not the first—to do so, and therefore suggests future direction for methodologically studying how individuals meaningfully produce and disseminate written language through mobile devices in real-time.

The following section will offer a theoretical framing for how text-based language may take on and perform social meaning, as well as review recent research that has to examined text through sociolinguistic, computational frameworks, and keystroke analysis. This background section will be followed by an overview of this study’s methods, followed by an overview of keystroke data collected, analysis of keystroke data in combination with discursive contexts in which those keystrokes occurred, and will conclude with a discussion that will suggest further study and directions for developing methods to examine text-based linguistic practices in mobile media.

THEORIZING AND TRACING TEXT-BASED LANGUAGE

This section will briefly provide a theoretical framework for studying text-based language in order to identify the exigence of this study’s methodological contributions. In doing so, I do not intend to provide an exhaustive review of literature in this area; rather, I intend to present a rationale for incorporating keystroke-logging methods to ask sociolinguistic questions about text-based performances through mobile media. I argue that it is useful to examine text-based language in communicative media through two underlying assumptions. The first assumption is that the composition and transmission of such text is part of individual social *performances* in individuals’ lived experiences. This performative conceptualization, which notably draws upon Goffman (1956), Butler (1997), and Coupland (2001), frames language use as continually constructing social identities, and each performance may contribute to meaning-making. Under this assumption, composing and sending text-based language through mobile media reflects how individuals perceive their own social identities and how they want to be perceived by their interlocutors.

The second assumption is that linguistic units transcribed into text can serve as material forms of *symbolic capital* that can signal social meaning and value when used as part of everyday practices (Bourdieu, 1984). For example, Eckert (2001) has argued that as adolescents navigate competing ideas of peer groups, adults, and their wider communities, they are “mutually engaged in the production of new meaning” (p. 34–35) through use of linguistic resources that form symbolic capital to *style* themselves according to one type of identity or another. Eckert (2008) further argued that through the very trading in their use, symbolic capital—linguistic or otherwise—may become associated with a variety of different possible social meanings, which Eckert terms the *indexical field*. Androutsopoulos (2014b) additionally argues that once a linguistic performance is recorded in audio-, visual-, or text-based media, that it is available material that individuals may use as stand-ins for symbolic capital, what he calls *media fragments*. These *media fragments* may become highly symbolic to trade in as part of performing social identities for particular peer groups (Georgakopoulou, 2014). Indeed, just as Eckert (2001) argued that clothing may signal one social identity or another among adolescent peer-groups, exchanging media fragments through frequent text-messages (Ling and Yttri, 2001; Laursen, 2005), emoticons or emojis (Baron and Ling, 2011; Highfield and Leaver, 2016), Facebook comments (Androutsopoulos, 2014a; Androutsopoulos, 2014c), or even YouTube videos (Georgakopoulou, 2014) may all serve to signal social identities, relationships, and even meanings that are continually being performed and negotiated.

These two assumptions frame text-based language-use as symbolic capital for performing social meaning, which may therefore more contribute to language change, for example, as media fragments are sent and received through various forms of media. Coupland (2001) has argued that in purposefully performing one style or another through broadcast media, such as radio, may in fact *produce* a shift in that meaning as they play with and balance perceptions of their audience and themselves. When these performances are received by audiences, it may change the indexical field of meanings attached to linguistic features because said linguistic features as well as their social meanings are diffused through broadcast media simultaneously. Coupland (2007) calls this process of new social and linguistic meaning developing through diffused broadcast media *decontextualization*, and Androutsopoulos (2014b) argues that decontextualized language may take on a range of new potential meanings that were construed through the broadcast and the audience's preconceived notions, a process he calls *recontextualization*. According to Androutsopoulos (2014b) recontextualized language reconfigures the ideological linkages in the indexical field of meaning.

This process of de- and re-contextualization may be readily observable through text-dominant mobile media, for example the rapid diffusion of *(on) fleek* on Vine and then Twitter (Grieve et al., 2017), and may be evidenced, as with spoken language, at both the social and individual level. In other words, the occurrence and prevalence of linguistic features might reflect broader social meaning, but the manner in which it is composed in real-time by individuals may reflect how, to paraphrase Coupland (2001), the individual is performed through the social. As I suggest later, understanding the processes through which individuals compose linguistic units into text may therefore be indicative of how they function to perform social meaning. The following section will therefore discuss how the production of text-based linguistic units may be seen as being performed in real-time, how paralinguistic cues, such as *Haha* and *Lol,* may evidence these performances, and how keystroke logging methods may aid in better understanding these performances.

Text and Time

Since the start of the new millennium various researchers at the intersection of sociolinguistics and media studies have documented the various social functions of text-based language, and how spoken communication practices are adopted for text-based media (see Ling, 2008). For example, researchers have observed how individuals use frequent text messaging to maintain contact (Ling and Yttri, 2001; Laursen, 2005), reconstitute paralinguistic meaning (Thurlow and Brown, 2003), adopt politeness strategies (Spilioti, 2011), circulate multimedia (Georgakopoulou, 2014), establish discursive structures and meanings for emojis (Sampietro, 2016; Pérez-Sabater, 2019), and even creatively play with spelling (Tagg et al., 2012; Tagg, 2012). Scholars examining social media have additionally examined how individuals may strategically modulate their audiences (Androutsopoulos, 2014a; Pavalanathan and Eisenstein, 2015a), perform social identities (Pavalanathan and Eisenstein, 2015b; Jones, 2015; Brock, 2020), pair with audio-visual channels (Piwek and Joinson, 2015; Highfield and Leaver, 2016and), or even participate in social media trends (Grieve et al., 2017). It is therefore noteworthy that scholars in the last 2 decades have identified—and continue to identify—the numerous ways that text-based language may be linked to different social meanings and functions that are continually negotiated through the technological affordances of the media through which text is circulated. An important example of this is the practice of what Thurlow and Brown (2003) term *paralinguistic restitution*, wherein individuals actively use the material affordances of the medium in order to communicate paralinguistic information, such as emotion or tone, that may be otherwise transmitted via prosodic features in spoken communication. In text-messaging, examples of paralinguistic restitution can include capitalization or reduplication to indicate stress (Thurlow and Brown, 2003; Fuchs et al., 2019), emoticons or emojis to indicate gesture (McCulloch, 2019), and *Haha* or *Lol* to indicate *shared laughter* (Jefferson, 1979; Ling and Yttri, 2001; Baron and Ling, 2011).

Further, corpus-based and computational methods, particularly for examining data from Twitter or text-messaging, have provided the opportunity to quantitatively aggregate and analyze text-based data in relation to broader sociolinguistic variables. As demonstrated by Eisenstein et al. (2014), Nguyen et al. (2015), Jones (2015), and Pavalanathan and Eisenstein (2015a), among others, how Tweeters tweet may reflect the configuration of their online social networks as well as regional demographic configurations of the geographic area from where they tweet (Jurgens, 2013; Eisenstein et al., 2014; Pavalanathan and Eisenstein, 2015b). In other words, what Tweeters tweet may reflect traditional "real-world" networks, communities, and online social identities.

While it is beyond the scope of this paper to fully survey and summarize the breadth of recent research in the emerging discipline called "computational sociolinguistics" (Nguyen et al., 2016), I do wish to echo and examine questions raised by Nguyen et al. (2016) that drive at the heart of linguistic research: *how do we locate individual agency in such data?* In their exhaustive survey of emerging computational sociolinguistic research over the last decade, Nguyen et al. (2016) argues that a central challenge of the nascent discipline is to reconcile the macro-scale informational and structural dimensions yielded from corpus research of text-based linguistic data with the real-world social performances and decisions of a flesh-and-blood person. Indeed, Nguyen et al. (2016) suggest that, in addition to examining traditional sociolinguistic variables, researchers can strive to further locate individual agency by incorporating "multimodal data" (p. 575). I suggest that such possible data modes and channels could detail how text is inscribed into a medium in real-time, as such data would illuminate the social, cognitive, and physiological processes through which text-based language is *articulated*. This form of *digital articulation* could include temporal and log-file about how individuals use their hands to input text to produce written language. Such data could complement computational methods to examine the widespread use of prevalent linguistic feature in Twitter and texting in order to better understand how those features are performed and meaningful to individuals.

Some researchers at the intersection of this field have indeed posited the connection between text and speech-based articulatory processes. Eisenstein (2013) and Eisenstein (2015) has examined numerous grapho-phonological "respellings" such as t/d-deletion or g-dropping for ing morphemes. Eisenstein (2015) argues that "when alternative spelling is linked to phonetic variation, it acquires at least the residue of the systems of phonological, grammatical, and social patterning present in speech" (p. 181); however, this may be dependent upon interactional contexts for addressing different audiences (see Pavalanathan and Eisenstein, 2015a), as syntactic and phonological constraints do not reliably predict variation in/ing and t/d-deletion in text as in speech. While Jones (2015) suggests that grapho-phonological variation on Twitter may reflect phonological realities of individual's speech patterns, I argue, as with Eisenstein (2015), that grapho-phonological variation in text-based linguistic performances may only reflect a "residue" of an individual's lived speech patterns rather than a verbatim transcription. The complex process of producing written language is drastically different from spoken language, as the interface of cognitive and articulatory processes involves overlapping *but materially different* physiological, psychomotor, interactional, and cognitive processes. Writing takes more time, requires psychomotor processes to control technological transcription, and results in asynchronous symbolic material. How individuals may compose, evaluate, or even strategically stylize text to satisfy their communicative needs may therefore be indicative of metalinguistic processes, and examining these processes may require theoretical and methodological consideration of how texts are produced. This study therefore follows Eisenstein's (2015) suggestion to use insights from keystroke-logging methods in order to examine how language is *digitally* articulated in real-time, and will do so through an investigation of two paralinguistic features: *Haha* and *Lol*. These features will be discussed in further detail in the next section.

Paralinguistic Features in Mobile Communication: Or, Who Laughs for Thee?

As noted above, individuals communicating through text-based media, such as text-messaging, may creatively use the affordances of the medium to engage in *paralinguistic restitution* in order to communicate information about paralinguistic features such as tone, stress, etc. (Thurlow and Brown, 2003). Spilioti (2011) offers a vivid example of *paralinguistic restitution* through analysis of closings in text-messages (e.g., *bye!* or *xoxo*), observing that, while closings may be typically absent from text-messages, the presence of closings may serve to strategically mitigate perceived face-threats via Brown and Levinson's (1987) politeness framework and signal relational closeness. The use of closings to communicate paralinguistic information is therefore predicated upon its expected norms (i.e., absence) related to everyday texting among interlocutors, which suggests that use of textual linguistic features may reconstitute paralinguistic meaning based upon frequency of use as well as how they are used within a text message. This study will examine two such paralinguistic features, *Haha* and *Lol*, and the remainder of this section will provide a brief review of these features and justify their selection for analysis in this study.

Haha and *Lol* may be seen as text-based representations of reacting with humor: *Haha* is a grapho-phonological approximation of laughter that has been recorded as far back as 1000 in Ælfric's *Grammatik und Glossar* (Ælfricof and Zupitza, 1880); while *Lol* is an acronym standing for "laugh out loud" that was possibly first coined in English-speaking internet chatrooms from the 1980s (McCulloch, 2019). Both features have been well-documented in digitally-mediated communication (DMC) research, for example, in computer-based writing registers such as Instant Messaging (Baron, 2004; Lewis and Fabos, 2005; Tagliamonte and Denis, 2008; Haas et al., 2011), and may be textually realized in a number of ways that conform to other common processes, such as reduplication or capitalization for emphasis (e.g., *hahahaha* or *LOLOL*). Regardless of whether *Haha* and *Lol* may be considered different variants of the same variable or different variables entirely (see Tagliamonte and Denis, 2008; Tagliamonte, 2016), both features appear to have maintained relatively stable usage across multiple studies over the last 2 decades. For example, Baron (2004) observed *Lol* made up 0.6% of all words in a corpus of IMs, Tagliomonte and Denis's (2008) found that variants of *Lol* and *Haha* made up, respectively, 0.41% and 1.47% of all words in a corpus of IMs, and Tagliamonte, 2016 similarly found that variants *Lol* and *Haha* made up, respectively, 0.69% and 0.40% of all word units in a corpus of IMs, text-messages, and other e-messages. As suggested by Tagliomonte (2015), these features have become present across other DMC registers, particularly *mobile* registers such as texting (Thurlow and Brown, 2003; Laursen, 2005; Baron and Ling, 2011) and even social media like Twitter (Pavalanathan and Eisenstein, 2015a), which is commonly interfaced through mobile devices (Brock, 2020).

The frequency and widespread usage across registers therefore suggests that *Haha* and *Lol* are relatively established and stable paralinguistic features in various registers of DMC (at least in the Western, English-centric world). Indeed, as will be further detailed below, participants in this study used these features with similar relative frequencies as noted above. I therefore argue that, because of the frequency with which these paralinguistic features are used, they are ideal to examine the usefulness of using keystroke-logging analysis to ask sociolinguistic questions. After all, the more frequently particular features are used by individuals on an everyday basis, the more behavioral keystroke data can be analyzed to examine articulatory patterns (e.g., how fast a feature is composed).

In addition to *Haha* and *Lol* being established paralinguistic features in texting, I suggest that they are also useful for examining interactional and situational contexts of texting. Thurlow and Brown's (2003) study demonstrated that texting conversations tended to manage relational intimacy as well as coordinate social discourse and activity. Paralinguistic restitution, which includes use of paralinguistic features such as *Haha* and *Lol,* therefore serves as part of these broader interpersonal functions, and this may be further seen through the ways in which texters frequently use such paralinguistic features to continually validate their relationship to one another (Ling and Yttri, 2001; Laursen, 2005). Through this lens, *Haha* and *Lol* do not just represent literal laughter, but may, as Baron (2004) suggested, structurally serve as "phatic fillers for the equivalent of OK, cool, or yeah" (p. 411). Further, studies of various paralinguistic features in texting (Ling and Yttri, 2001; Highfield and Leaver, 2016; Sampietro, 2016; Pérez-Sabater, 2019) have commonly observed that such features occur at turn-taking boundaries (i.e., the start or end of a text message), and, to a lesser extent, clause boundaries within a message. This suggests that *Haha* and *Lol* in texting might serve multiple intersecting functions: as symbolic capital to manage interpersonal relationships, and to coordinate turn-taking structures similar to *shared laughter* (Jefferson, 1979). From this vantage point, *Haha* and *Lol* indeed maintain some of the "residue" of spoken language as Eisenstein (2015) suggests, but also function according to the ways in which individuals negotiate the technological affordances and discursive expectations of text-messaging registers.

I further suggest, following Spilioti (2011), that Brown and Levinson's (1987) politeness framework is powerful for sociolinguistic interpretations of the specific interactional contexts in which *Haha* and *Lol* may be used in texting. Within Brown and Levinson's (1987) framework, individuals may use various politeness strategies during communication in order to mitigate possible face-threatening acts (FTAs) to themselves and interlocutors. According to Brown and Levinson (1987), FTAs can occur and be mitigated through linguistic, non-linguistic, or paralinguistic channels, and can affect an individual's *positive face*, i.e., "the desire [. . .] to be approved of" (p. 13), or *negative face*, i.e., "the desire to be unimpeded in one's actions" (p. 13). For example, an individual may use *shared laughter* via a "laugh particle" (Jefferson, 1979) at the start of a turn in response to an interlocutor's joke in order to avoid damage to the interlocutor's positive face (i.e., in order to preserve the interlocutor's self-value of being humorous and liked); while laughter at the end of a turn may preserve the positive face of the speaker and mitigate negative face-threats to the interlocutor (i.e., in order to preserve the speaker's self-value and avoid imposing upon the interlocutor). As Spilioti (2011) suggested, since texting conversations and the asynchronous turn-taking structure serve to manage relational work, Brown and Levinson's theories frame every sent and received text message as symbolically imbued with politeness strategies. For example, since individuals are compelled to send messages frequently (Ling and Yttri, 2001) and respond to messages quickly (Laursen, 2005) to signal relational closeness, texting conversations may be seen as continually navigating politeness strategies because every sent message is a negative FTA (i.e., because it imposes on the receiver) and every response is a positive FTA (i.e, because it signals how the receiver is valued). Turn-taking positions in a text message (i.e., the start or end of a message) may therefore be seen as highly salient positions through which texters work to mitigate such FTAs, and the use of *Haha* or *Lol* at these turn-taking position may symbolically negotiate these politeness strategies.

Further, as noted by Meredith and Stokoe (2014), asynchronous text-based channels such as texting and IM afford individuals the ability to manage and even repair execution of these politeness strategies in the message space unseen by the interlocutor, i.e., prior to sending the message. Meredith and Stokoe (2014) found that such *message construction* repairs bear similarity to repair work in spoken language, albeit while remaining unseen and therefore "unaccountable for some interactional matters" (p. 202). This suggests that the seemingly unseen processes through which individuals cognitively select and compose specific textual features reflects how individuals strategically manage politeness strategies. The timing processes for composing *Haha* or *Lol* at different turn-taking positions, which are seen only by the individuals composing the message (Meredith and Stokoe, 2014), may reflect the *residue* of how individuals are cognitively processing these strategies in order to manage relationships with their interlocutors. This requires attention not only to how and where *haha* and *lol* are distributed in sent and received messages—which may be accomplished through text-based linguistic analysis—as well as the timing processes through which these features are composed—which may be accomplished through keystroke-logging analysis.

Keystroke-Logging Analysis and Digital Articulation

In this section I will provide a brief contextualization of keystroke-logging analysis and how it may offer articulatory evidence for the linguistic production of text. Keystroke-logging analysis has roots in writing studies, an area of research that emerged in the 1980s and has pulled in researchers from various fields, such as cognitive science, applied linguistics, educational psychology, technical communication, etc (see Hayes, 1996; Cislaru, 2015). Central to this methodology is examining temporal data from writing in order to infer cognitive processes that are engaged *during* writing (Plane, 2015). This requires examining a textual artifact according to how it was composed in real-time, i.e. *synchronously*, in conjunction with the completed text as the primary source for extricating linguistic meaning. I use the term *digital articulation* purposefully, as the metaphor of articulation—which linguists often think of in physiological, perceptual, and acoustic terms—in order to call attention to the fact that composing linguistic material in text involves an articulatory process that unfolds synchronously (Farman, 2012; Plane, 2015). This bears some similarity to speech, except that writing involves digital[1] articulatory mechanisms and results in an asynchronous textual artifact. This differentiation from speech therefore requires writing researchers to unpack what Grésillon and Perrin (2015) term the *double black box*, i.e., the processes through which a text was composed as well as how those processes evidence cognitive and social processes involved in composition.

Keystroke logging therefore serves as one such methodology to unpack this *double black box*, as it allows analysis of temporal records of discrete input-based events involved in writing with a computer, i.e., pressing individual keys on a keyboard in order to compose a digital text. Leijten and Van Waes (2013) argue that keystroke logging allows researchers to both re-construct the temporal processes through which individuals composed a text *and* to observe writers rather unobtrusively. Further, because keystroke logging's primary unit of analysis is discrete keystroke events, temporal analysis is located primarily in *pauses*, i.e., the time in *between* the input of individual keystrokes. This conceptualization of the *pause* borrows heavily from speech production, in which, as argued by Miller (2006), pausing during writing provides indirect and inferential evidence of writers' cognitive resources, including attention management and long-term memory retrieval.

Pause-based data has therefore been shown to relate to linguistic characteristics (Van Waes, Leijten, Lindgren, and Wengelin, 2015). This requires coordinating pause-based analysis, typically through the measurement of time between key presses called the *inter-key interval* (Leijten and Van Waes, 2013), as well as the processual sequence of keystrokes that construct recognizable linguistic units of information. For example, a given sequence of keystrokes, such as [H], [e], [l], [l], [o], therefore may represent the intended construction of the word *Hello*, and the inter-key intervals for the first key may therefore be longer than the intervals for all subsequent keys. Using keystroke analysis to examine linguistic content is therefore, in some ways, similar to using acoustic analysis in order to examine phonological variables such as speech rate (see Kendall, 2013), as pause-times will often distinguish between smaller and larger chunks of linguistic units. For example, the pause between [e] and [l] in *Hello* will be shorter than between [o] and the first letter of the next word (Van Waes et al., 2015).

Researchers have therefore taken multiple approaches in order to examine linguistic segments, particularly latencies between different syntactic units, morpho-phonological syllable boundaries (see Nottbusch, 2010), as well as use specific keystroke events and pause times in order to distinguish boundaries between *pause bursts* (i.e., keystroke activity between pauses over 2000 milliseconds) or *revision bursts* (i.e., keystroke activity relating to revising text) (see Galbraith and Baaijen, 2019). For example, using *Inputlog*[2], Leijten et al. (2012) incorporated various NLP tools on linear notation of keystroke data (called S-notation) that represents the non-linear process of composing textual products. Leijten et al. (2012) argues that this allows analysis of word-level revisions (i.e., individual words that are revised for individual characters), deleted segments (i.e., multi-word units that are deleted within the same sequence of deletion activity), and the final text; all of which are subject to part-of-speech (PoS) tagging, lemmatization, chunking, and word frequency analysis. In addition, Olive and Cislaru (2015) combined both *Inputlog's* NLP analysis to compare with corpus-based methods to examine the timing processes of *pause bursts* as well as *repeated segments* (i.e., a sequence of two or more linguistic units that occur at least twice in a corpus), and found evidence that only 3% of these units shared overlapping syntactic structures. Further, in examining text produced by college students taking a test ($N = 38$), Plank (2016) additionally found that, when applying a Long-Short Term Memory model (a type of bi-directional neural network), pauses helped illuminate chunking at the word-level, but not necessarily the morpho-syntactic level. This research demonstrates that keystroke-log data may be organized in a number of ways in order to analyze recognizable linguistic units of information, particularly temporal analysis surrounding word boundaries.

Further, considering the idiosyncratic nature of individual writers (Plank, 2016) and variations across written registers individuals may be familiar with/have access to (Biber and Conrad, 2009), keystroke analysis has demonstrated the value in looking more qualitatively at individual writers in order to more robustly examine how text is produced in context. For example, in examining writers with dementia, Leijten et al. (2015) found that such writers required much more time than non-dementia writers of similar ages to compose nouns and verbs compared to articles or prepositions, which the authors suggest reflects the greater cognitive demands placed on writers with dementia. Leijten et al. (2019) applied similar methods to compare native Dutch speakers writing in Dutch (L1) and English (L2), observing that pause-based differences may be attributed to language, word length, and PoS, and that these pause patterns may repeat for frequent two- and three-word constituents. Importantly, Leijten et al. (2019) also observed that, based on pause-based analysis, language differences primarily were limited to spelling and word choices. Leijten et al. (2014) additionally demonstrated that examining even a single writer producing a single text over a several-day period may yield important theoretical insights about how writers may use schematic knowledges of various genres and registers to construct texts. Most recently, Bowen and Van Waes (2020) used keystroke logging and systemic functional linguistics frameworks in order to examine revisions during writing,

[1] I use the term *digital* to simultaneously reference the fact that writing is often composed through both computational technologies as well as inputted via digits on human hands.

[2] Developed by Leijten & Van Waes (2013).

including the finding that revisions may most frequently occur at or just before the point of inscription. While Bowen and Van Waes (2020) also only observed a single writer composing over multiple writing sessions, their study demonstrated the rich possibilities of applying keystroke analysis in order to examine linguistic frameworks. Indeed, the amount of data obtained via keystroke logging from individual writers over longer spans of time, rather unobtrusively and indirectly, allows for in-depth analysis and consideration of how individuals may meaningfully construct written language in context in order to contribute to theory-building to conduct broader and more generalizable studies.

While only a short sample, these examples demonstrate both the value of using keystroke analysis to ask linguistic questions regarding text production. These studies may privilege lexical and morpho-syntactic analysis, partially due to the incorporations of NLP methods; however, as suggested by Bowen and Van Waes (2020), Eisenstein (2015), and Nottbusch (2010), keystroke analysis remains promising for asking sociolinguistic questions. For example, *how might the timing processes of enregistered and unmarked sociolinguistic variables differ*? *How might they differ for variables that are undergoing a change-in-progress through rapid diffusion across social networks*? *Would any such differences indicate how individuals and social networks recontextualize variables differently*? Consider the rapid diffusion of *(on) fleek* through Twitter in 2014 (see Grieve et al., 2017), which may be attributed to a viral video. An examination of the timing processes through which Twitter users composed *(on) fleek* in order to contribute to its rapid diffusion may provide insights into how this feature was adopted by users. For example, as suggested by keystroke analysis literature, would newly adopted linguistic features, those that are highly salient, or those undergoing a change to their indexical field of meaning, be composed more slowly or experience longer pauses before and after inputting? Keystroke analysis, in addition to analysis of the frequency of use and discursive structuring of these text-based features, could therefore illuminate how individuals are meaningfully taking part and contributing to language change.

Lastly, I suggest that in order to examine the wealth of text-based data that may be commonly diffused through social media, it is important to expand beyond computer terminals. After all, well over 80% of Twitter use may be conducted on mobile devices (WSJTech, 2014, Apr 14), and writing interfaces on mobile touchscreen devices involve qualitatively different input-processes from other writing interfaces (see Farman, 2012; Mangen, Anda, Oxborough and Brønnick, 2015; Parisi, 2018). Indeed, as suggested by Brock (2020), Twitter borrowed heavily from SMS architecture and interfaces, and therefore likely encouraged compositional habits similar to texting. Therefore, even though preliminary studies into writing processes for composing *simulated* tweets on computers has demonstrated value (see Leijten et al., 2012), observing the compositional processes of these registers as they occur on mobile devices may yield more “naturalistic” observations.

Research Questions

Based on the above, this study therefore seeks to examine the following research questions based on the keystroke-logging data:

RQ1: What are the frequencies of occurrence of paralinguistic variables *Haha* and *Lol* in text messaging, and what is their distribution according to turn-taking structures in asynchronous messaging?

RQ2: What are the timing processes of *Haha* and *Lol* in text messaging, and how do these timing processes reflect turn-taking structures in asynchronous messaging?

As will be detailed in the following section, this study therefore applies keystroke-logging methods for writing on mobile devices, and further explores further means of asking sociolinguistic questions for text-based language that is common to written registers in mobile communication.

METHODS

The present study reports on keystroke-logging of mobile devices. This follow Eisenstein’s (2015) suggestion that keystroke-logging methods may more closely observe the production of text-based linguistic content in popular mobile platforms, as well as Schneier and Kudenov’s (2018) demonstration of how keystroke-logging data can be successfully collected from mobile devices. This study was therefore designed to observe how text is digitally articulated on mobile devices as individuals compose and send text-messages to members of their social network. This involved developing a mobile keyboard application for Android devices to log keystroke data, collecting keystroke data from participants ($N = 10$), and conducting pause-based analysis of keystroke data pertaining to paralinguistic features from seven of those participants. This section will provide further details on the designs of LogKey, how data was collected as part of this study, briefly overview data output and analysis, and discuss this study’s sample size.

Designing LogKey

In the recent decade, writing scholars, particularly Van Waes et al., (2012) and, have made concerted efforts to establish standardized recommendations for designing keystroke loggers for computer-based writing. These recommendations outline use of XML-structure to log the sources of computer-based actions, such as input from a keyboard or mouse, as well as how to operationally (or even algorithmically) define a sequence of actions, such as how to define the temporal threshold of a *pause* during writing. While Van Waes et al. (2012) do discuss methods of how to accommodate other means of input, particularly through speech recognition, use of a stylus, or even use of the “swipe” action on a touchscreen device, these recommendations do not explicitly address how a keystroke on a computer keyboard with tactile keys is *not* the same as a keystroke on a virtual keyboard.

As found in Schneier and Kudenov (2018), the technological distinctions between a computer keyboard and a virtual keyboard have a significant impact for how to log keystrokes. Computer keyboards have keys with binary up or down depressions; while virtual keyboards have keys that *simulate* the up or down depression of a key according to how the electronic charge of a finger comes into contact with the corresponding image of a key

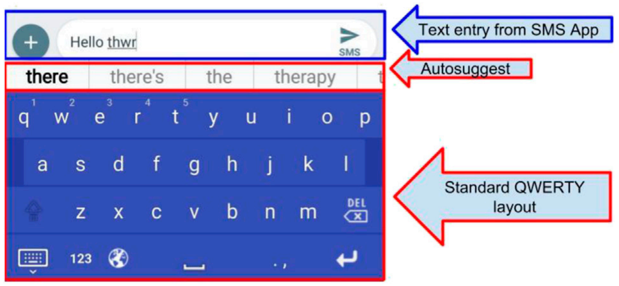


FIGURE 1 | The LogKey keyboard, with autosuggest options.

on screen (Andre et al., 2005; Westerman and Elias, 2006). Van Waes et al. (2012) outline that the times of each down press of a key and release of a key should be measured in order to determine the time between one key's release and the next key's depression, what is called the inter-key interval (IKI). On a touchscreen, though, how can we measure the IKI if there is no *depression* of a key but instead *contact* with the screen?

The method for logging keystrokes in this study therefore intended to accommodate the technological configuration of what it means to press a virtual key on a virtual keyboard, as well as improve upon the methods explored in Schneier and Kudenov (2018) wherein participants ($N = 5$) used a smartphone with a keystroke logger built directly into the functionality of this phone. The method for keystroke logging in this study therefore intended to, 1) allow participants to use their own personal mobile devices in order to observe them using devices they were presumably familiar with and comfortable using; 2) observe participants over a longer span of time in order document everyday compositional habits on their mobile devices; 3) log what applications individuals were using when keystrokes were logged; and 4) log the time between the initial press of the key that was pressed and the previous key. In regards to item 4, this method replicates Schneier and Kudenov's (2018) operationalization of the IKI, which Plank (2016) suggested is most valuable. In order to address the above needs, a virtual keyboard application was designed and constructed for this study, an application which could be substituted for the standard virtual keyboard. This app, called LogKey, was designed for the Android mobile operating system[3].

The primary features of LogKey, from the perspective of users interacting with the Graphical User Interface (GUI), is that the application would appear as a standard QWERTY layout virtual keyboard, and include autosuggested text (see **Figure 1**). Like the standard Android keyboard, this virtual keyboard acts as a separate layer on top of whichever application is in use.

Study Procedures

Participants ($N = 10$) were recruited through snowball sampling methods, primarily through various online message-boards and list-servs commonly accessed by undergraduate and graduate students at a large university in the mid-Atlantic region of the United States. Recruitment materials informed participants that the study was intended to learn about how individuals communicate through text-messaging, and would involve using an unreleased keyboard application for Android OS. Sociolinguistic studies of mobile communication, particularly texting, frequently involve university students (Thurlow and Brown, 2003; Baron and Ling, 2011; Spilioti, 2011; Androutsopoulos, 2014a) because they tend to be technologically verbose (Baym et al., 2004; Ledbetter, 2008; Lenhart, 2015) and may be seen as transitioning between youth- and adult-centered identity practices (Eckert, 2000). Further, Pew Research Center reports from 2014 further suggest that they (by the time of the target date of the study in 2018) are accustomed to predominantly using mobile smartphones (76% of 15–17 year-olds) for a range of purposes including general internet use, text-messaging, video chat, social media use, other messaging applications, and various other communicative applications (Lenhart, 2015).

Following completion of the Informed Consent process, participants were asked to complete two interviews (pre- and post-study), a brief observation session, and use the LogKey keyboard application for a period of seven days. The 5–10-min pre-study interview briefly addresses participants' habits and history using mobile communication media, which social ties they generally communicate with and how they describe those relationships (e.g., close friend, parent, roommate), as well as participants' perceptions and attitudes regarding their mobile communicative practices within their interpersonal networks (Ledbetter and Mazer, 2014). Following the pre-study interview, participants were instructed how to download and install the LogKey application, as well as obtain Third-Party Consent through an Interlocutor Consent Script[4] that was sent to individuals they expected to text regularly throughout the week, as well as any individuals who would text them later in the week. Obtaining third-party consent was deemed important because even though these third parties were not directly participating in the study (i.e., they were not using LogKey to have their keystroke data logged), the messages that they sent to participants and the messages that participants sent back to them (both of which were part of data collection) arguably *belonged to both* parties, and contained private information regarding both parties meaning these third parties had the ethical right to consent to participate indirectly or not.

[3] LogKey was designed and implemented by Nicholas Miano over several months while in consultation with the author. Dr. William Enck, Associate Professor of Computer Science at North Carolina State University, conducted code review to ensure that the app met stringent security standards in addition to a review by university IRB.

[4] The Informed Consent Script informed participants' interlocutor that the participant was involved in a study that would be collecting their keystroke data and downloading SMS conversations from their phones, if the interlocutor consented in writing as well. The script further explained that if the interlocutor consented, messages sent from them would be used to contextualize content that the participant composed.

TABLE 1 | Example of log file data in comma-separated value form.

Date	Event Log	Event Type	Time MS	Pause MS	Application
14:30:48	t	Key	1505154648754	553	Messaging
14:30:48	h	Key	1505154648918	68	Messaging
14:30:49	[The, that, this, they, there]	Auto_options	1505154649008	n/a	Messaging
14:30:49	e	Key	1505154649086	168	Messaging
14:30:49	SPACE	Key	1505154649679	549	Messaging
14:30:50	e	Key	1505154650478	696	Messaging
14:30:50	m	Key	1505154650630	55	Messaging
14:30:50	o	Key	1505154650866	128	Messaging
14:30:51	j	Key	1505154651241	281	Messaging
14:30:51	[Emoji, emojis, emojis, Emoji's]	Auto_options	1505154651306	n/a	Messaging
14:30:51	i	Key	1505154651422	153	Messaging

At the conclusion of the seven days participants met again with the researcher. During this meeting the researcher instructed participants on how to securely transfer keystroke log-file data from the LogKey application to the researcher. Participants were also instructed how to download and install the SMS to Text application (SMeiTi, 2017), and to export textual log-file data from application to their SD card, and then transfer electronically to the researcher. The researcher then went through the log-file data with the participants in order to conduct a spot-check of the data, particularly to make sure that no data from third-parties who did not grant written permission to the participants be included in the data set. Following this, the researcher then conducted a brief post-study interview that discussed the participants' experiences using the LogKey application throughout the week as well as what conversations they engaged in through their mobile devices.

Keystroke Data Output

Adapting the recommendations from Van Waes et al. (2012), as well as lessons from Schneier and Kudenov (2018), keystrokes from use of LogKey were logged and stored in a comma-separated value (CSV) file format, as demonstrated in **Table 1**. This log file separates each log event into individual rows and is sorted chronologically, and each log event then has several corresponding values expressed in individual columns including: The date and time of the event (Date); The value of the log event as seen from the keyboard layout (Event Log); The category of the log event (Event Type); The Unix time in milliseconds of the initial press of the key (Time MS); The time elapsed between the initial press of the log event and the previous log event's initial press (Pause MS); The application in use when the log event was recorded (Application). As can be seen in **Table 1**, this output can show when and what autosuggest options are presented to the user, as well as when and what autosuggest option they choose.

A disadvantage to the software configurations of mobile keyboard application on Android is that, as mentioned, only the simulated keys of the keyboard itself can be logged. This means that depending upon the application that the keyboard is being used with, such as the various SMS applications for Android, transmitting a message may not be recorded as a log event. In addition to not knowing when a text message is sent, LogKey is also unable to present a summative record of the message that was sent to interlocutors. The application SMS to Text, however, allowed participants to export SMS messages to the SD card on their personal mobile device. These exported messages may be saved in the form of a text (.txt) file or comma-separated value (.csv) file., and included information about the time of transmission, whether the message was sent to or from the participant, the number of the interlocutor, the name stored in the participant's contacts list, and the textual content of the message sent or received. Together with the keystroke log-file data, as seen in **Table 1**, the text logs can be coordinated so that keystrokes may be corresponded to specific messages.

In merging and coordinating both data sources into a single matrix, it was possible to examine how a synchronous sequence of keystrokes constructed an asynchronous message that was transmitted from a participant's device. Further, using a combination of computational and manual coding, it was possible to demarcate individual word units as well as entire message units from a given sequence of keystroke activity. Using LogKey's data output that displayed each sequential keystroke, the associated alphanumeric key or the SPACE key, and the IKI of that key, it was possible to identify series of sequential alphanumeric keys that may represent the intent to type a particular word. For example, an [H] key with an IKI above 2000 ms, followed by the keys [e], [l], [l], and [o] followed by the [SPACE] key likely represents a sequential burst of alphanumeric keystrokes activity from typing the word 'Hello' that I term *keybursts* or KBs. A series of embedded If-Then conditional formulas in MS Excel then labeled the first alphanumeric key in a keyburst (i.e., the inter-KBs, or the pause before the keyburst) that occurred immediately after non-alphanumeric keys or keybursts over 2000 ms, as well as the following alphanumeric keys in the keyburst up until a non-alphanumeric key (i.e., the intra-KBs). Doing so made it possible to group all keys within a keyburst and to distinguish the timing processes of the first key in a keyburst from the others (e.g., the [H] would be labeled "inter-key" and the keys [e], [l], [l], and [o] each labeled 'intra-key'). Future versions of LogKey may designed to computationally produce this data as log-files are compiled, rather than through post-hoc tagging.

As discussed above, chunking keystroke activity into bursts, such as *pause* or *revision bursts*, is a common practice in keystroke analysis in order to infer linkages between linguistic, writing, and cognitive processes (Leijten et al., 2015; Galbraith and Baaijen, 2019). Further, boundaries for word units are often distinguished

TABLE 2 | the top 75 most frequent keybursts in texting data.

rank	Keyburst	Absolute frequency	Relative frequency	Characters
1	i	334	3.56%	1
2	To	224	2.39%	2
3	The	167	1.78%	3
4	You	139	1.48%	3
5	And	122	1.30%	3
6	Im	121	1.29%	2
7	a	119	1.27%	1
8	t	103	1.10%	1
9	is	90	0.96%	2
10	It	78	0.83%	2
11	me	73	0.78%	2
12	So	73	0.78%	2
13	That	69	0.74%	4
14	do	68	0.73%	2
15	my	65	0.69%	2
16	u	63	0.67%	1
17	e	61	0.65%	1
18	No	59	0.63%	2
19	s	59	0.63%	1
20	But	56	0.60%	3
21	y	54	0.58%	1
22	Have	53	0.57%	4
23	we	51	0.54%	2
24	d	50	0.53%	1
25	Of	50	0.53%	2
26	Just	49	0.52%	4
27	are	48	0.51%	3
28	For	48	0.51%	3
29	In	48	0.51%	2
30	o	48	0.51%	1
31	be	43	0.46%	2
32	n	43	0.46%	1
33	w	43	0.46%	1
34	At	42	0.45%	2
35	Good	38	0.41%	4
36	Oh	38	0.41%	2
37	On	38	0.41%	2
38	g	37	0.39%	1
39	Its	36	0.38%	3
40	Or	36	0.38%	2
41	was	36	0.38%	3
42	What	36	0.38%	4
43	All	35	0.37%	3
44	Dont	35	0.37%	4
45	How	35	0.37%	3
46	I	35	0.37%	1
47	not	35	0.37%	3
48	If	34	0.36%	2
49	m	32	0.34%	1
50	r	31	0.33%	1
51	Go	29	0.31%	2
52	h	29	0.31%	1
53	Get	28	0.30%	3
54	Will	28	0.30%	4
55	Yeah	28	0.30%	4
56	Your	28	0.30%	4
57	This	27	0.29%	4
58	Up	27	0.29%	2
59	can	26	0.28%	3
60	Lol	26	0.28%	3
61	Okay	26	0.28%	4
62	f	25	0.27%	1
63	k	25	0.27%	1
64	Like	24	0.26%	4
65	Hey	23	0.25%	3
66	I	23	0.25%	1
67	Want	23	0.25%	4
68	Know	22	0.23%	4
69	Love	22	0.23%	4
70	An	21	0.22%	2
71	b	21	0.22%	1
72	he	21	0.22%	2
73	Time	21	0.22%	4
74	Haha	20	0.21%	4
75	Now	20	0.21%	3

(Continued in next column)

TABLE 2 | (*Continued*) the top 75 most frequent keybursts in texting data.

by use of function keys, i.e., the space bar or keys for punctuation (Van Waes et al., 2015). In other words, analysis of keyburst units will be similar in form and function to identifying potential word units according to keystroke pauses tagged as "BEFORE WORD" or "AFTER WORD" in *InputLog* (Leijten and Van Waes, 2013; Van Waes et al., 2015).

Identifying sequences of keystrokes in such a manner is a highly productive method for operationalizing input-based activity with purposeful communicative practices (see Van Waes et al., 2012); albeit, it does make at least two assumptions. The first assumptions is that a given sequence of alphanumeric keystroke logs reflect a participant intentionally composing a particular linguistic unit. While this may seem rather matter-of-fact, as Suchman (1987) has cautioned, human input and interaction with machinic interfaces (i.e., a log file record) reflects the machine's designs and constraints as much as the human user's intentions. In other words, at the same time a given burst of keystrokes may reflect a participant intentionally writing a particular word, it may also reflect unintentional keypresses (i.e., a pocket-dial or pressing an errant key). A second assumption is that identifying keybursts, particularly the timing processes of these keybursts, will provide important evidence of psychomotor processes involved in word recall and procedurally inputting that word through QWERTY keyboard. Writing process research has suggested that the time between the first key inputted in a burst of keystroke activity or between typing separate words evidences such cognitive processes (Miller, 2006; Hayes, 2012; Leijten and Van Waes, 2013; Leijten et al., 2014; Mangen et al., 2015). This study will therefore focus largely on the time before a keyburst was inputted, i.e., the IKI value of the inter-KB key, and the average time required to input all subsequent keys in the keyburst, i.e., the average IKI value of the intra-KB keys in a keyburst.

Linguistic Analysis of Keystroke Data

In order to investigate the research questions, this study first identified all variants of *Haha* or *Lol* that occurred in 908 sent text-messages that participants procured for this study. This involved creating a matrix that identified each token, the message it occurred in, its turn-taking position (i.e., initial, medial, or terminal), and, by coordinating with the keystroke data, the inter-KB and intra-KB values. It should also be noted

TABLE 3 | Frequency and occurrence of all Haha or Lol Keybursts from Text-Messaging Data.

Variable	Variant	Occurrrence	Frequency of texting keybursts
Lol	Lol	26	0.2773%
Lol	LlolI	1	0.0107%
Lol	Lol	1	0.0107%
Lol	Lolim	1	0.0107%
Lol	Lolol	1	0.0107%
Lol	Lolthsy	1	0.0107%
	Total lol keybursts	**31**	**0.3307%**
Laughter	Haha	20	0.2133%
Laughter	ha	4	0.0427%
Laughter	Hah	2	0.0213%
Laughter	Hahahahaha	2	0.0,213%
Laughter	Ehe	1	0.0107%
Laughter	Ehehehe	1	0.0107%
Laughter	Haah	1	0.0107%
Laughter	Hahah	1	0.0107%
Laughter	Hahah	1	0.0107%
Laughter	HAHAH	1	0.0107%
Laughter	Hahaha	1	0.0107%
Laughter	Hahahah	1	0.0107%
Laughter	Hahahaha	1	0.0107%
Laughter	Hahahahahha	1	0.0107%
Laughter	Hahahahhh	1	0.0107%
Laughter	Hahahit	1	0.0107%
Laughter	Hwhw	1	0.0107%
Laughter	Jehe	1	0.0107%
	Total laughter	**42**	**0.4481%**

that the turn-taking position was determined by the position of the variable within the transmitted text-message, not within the sequential process of composing the message, and that the code for the initial position included instances in which a transmitted message only contained one of these variables.

Further, using the dplyr[5] package in R Studio, this study generated a matrix with every unique keyburst produced by all participants, including the number of times each keyburst occurred, and its relative frequency among all keybursts. Using this matrix, it was possible to hand-code all unique variants of *Haha* and *Lol* (see **Tables 2** and **3**), and then identify every occurrence of *Haha* and *Lol* throughout the keystroke data (texting and all) as well the application in use when each token was composed. Due to the manner in which keyburst boundaries were determined (see ***Section 3.3***), variants necessarily included keybursts that were never transmitted, likely a result of repairs related to typing errors. For example, as shown in **Table 3**, the list of variants of *Haha* include readily apparent variants such as *haha* and *hahah,* as well as unapparent variants such as *hwhw* and *jehe*. These two variants, which occurred only once each, were necessarily identified as variants of *Haha* because they were deleted and replaced with more recognizable *haha* or *hehe*, and because the repaired characters are adjacent to the mistyped characters on the QWERTY keyboard. I include such keybursts as variants in this list because, even though they are clearly not spelled the same as the more iconic and frequent variants and are most likely *typos*, these keybursts likely reflect habitualized articulatory processes the same as the others.

A Note About Participant Pool Size

Since this study included only 10 participants, its findings are not necessarily generalizable. Nevertheless, I wish to put this small sample size in the context of writing process research. First, as discussed above, keystroke-logging studies may often involve a small number of participants. For example, Leijten et al. (2014) involved $N = 1$, Leijten et al. (2015) involved $N = 2$, and Bowen and Van Waes involved an $N = 1$. In such studies, researchers are less concerned with generalizing about entire populations and more so focused on examining writers in-depth and in-context in order to challenge and build upon theoretical models of writing (see Leijten et al., 2014). Second, in spite of the small participant pool, this study observed individuals writing for longer durations and as part of everyday mobile communication habits. This compares drastically to common keystroke studies that writing tasks of shorter duration in formal settings. For example, participants in Leijten et al. (2012) composed short simulated tweets; participants in Leijten et al. (2015) composed texts typically written in under 10 min; participants in Van Waes et al. (2010) revised short sentences; and participants in Nottbusch (2010) composed short sentences in response to stimuli. Even in studies that observed writers composing formal reports over multiple days, such as Leijten et al. (2014) and Bowen and Van Waes (2020), participants engaged in individual writing episodes that would last between 20 min to several hours. In other words, a study of writing on mobile devices, which may involve shorter forms of writing, may nevertheless yield similar data sets as Leijten et al. (2014) and Bowen and Van Waes (2020), in addition to involving writing more frequently throughout a participant's everyday life.

RESULTS

This section will provide an overview of the keystroke data collected from this study. I will start with more descriptive summary of the data collected from participants' use of the LogKey keyboard across all applications, and then narrow in further by discussing the occurrence of *Haha* and *Lol* in the keystroke data in order to address RQ1, and the timing processes of those features in order to address RQ2.

General Overview

The participants in this study (N = 10) reflect the targeted population, in that they all used mobile phones running Android OS as their personal devices, were between the ages of 18–35, and were all college-educated. A majority of the participants were currently enrolled in undergraduate studies at a four-year institution ($N = 7$), while the remaining participants were either enrolled in graduate studies ($N = 2$) or were working

[5] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.5. https://CRAN.R-project.org/package=dplyr

TABLE 4 | General summary of keybursts for all participants.

Participants	Total keybursts	Total keybursts >1 keystroke	Median of inter-KB IKI	Median of average intra-KB IKI	Median of intra-KB IKI average (≥3 characters)	Median of intra-KB IKI average (<3 characters)
A	4,585	3,727	184	96	65.0	144
B	152	127	492	110.95	76.33	154.0
C	4,311	3,652	216	148.8	98.0	221.22
D	1,223	1,107	106	62	47	71.60
E	4,375	3,731	369	206.4	146.0	275.80
F	2,121	1786	216	86.0	52.00	151.5
G	3,119	2,716	229	111.1	78.3	158.5
H	6,353	5,377	104	54.0	41.0	67.80
I	5,071	4,381	193	117.5	95.0	140.80
J	964	863	183	171.6	99.0	217.0
Total	**32,274**	**27,467**	**201**	**103**	**72**	**149**

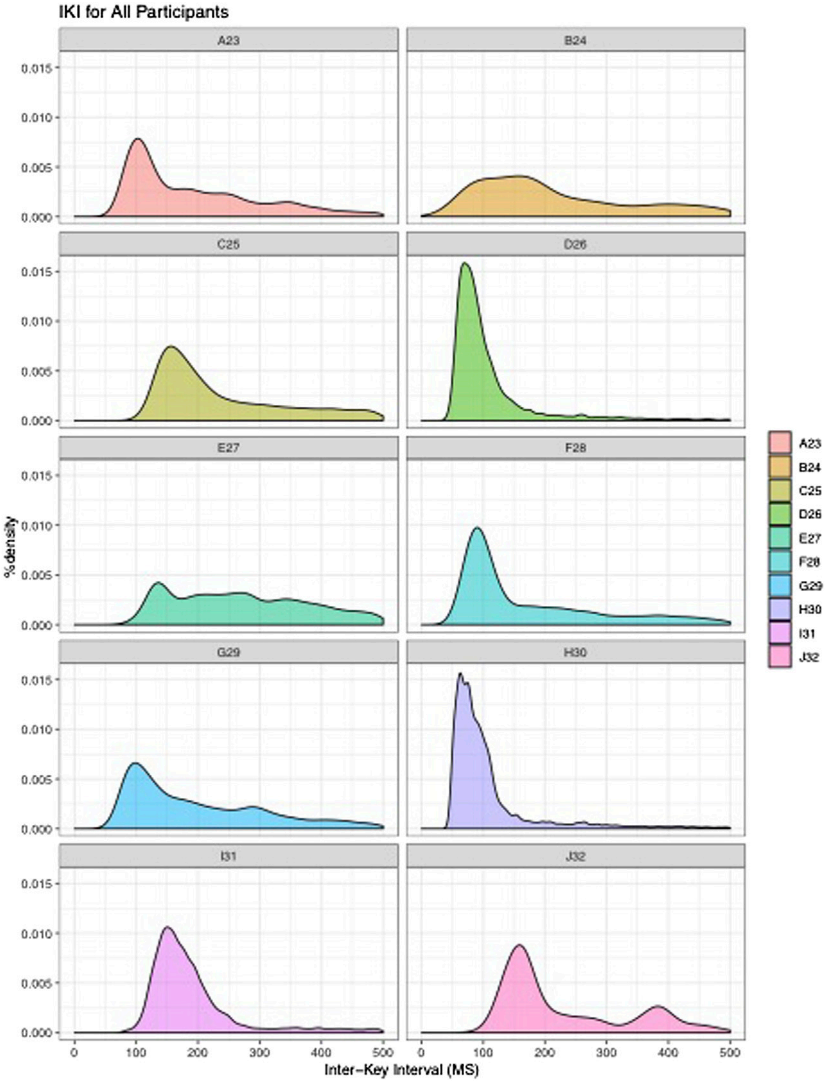


FIGURE 2 | Density plots showing IKI for all participants.

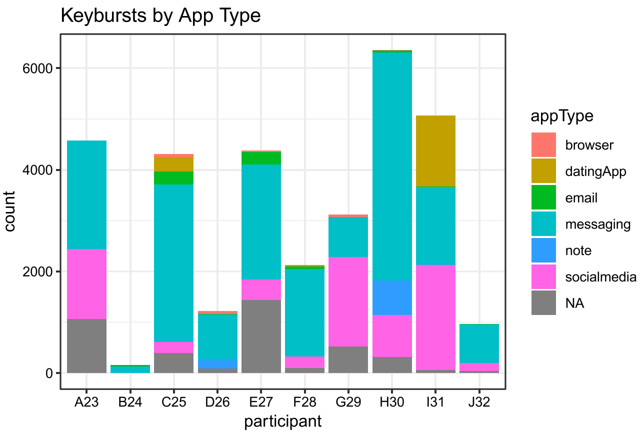


FIGURE 3 | Keyburst count for Application Type for each participant.

professionals (N = 1). All participants reported that text-messaging was among their most-used communication or messaging applications, although apps such as Snapchat, Instagram, GroupMe, WhatsApp, Facebook, Facebook Messenger, Twitter, and email were frequently used as well. In total, 179,996 keystrokes, or 32,274 keybursts (see **Table 4**) were collected from all 10 participants' use of LogKey, with the fewest number of keybursts produced by Participant B (N = 152), and the most produced by Participant H (N = 6,353). All participants varied from one another, although each participant displayed a general tendency for typical IKIs to cluster below 500 ms (see **Figure 2**). Nevertheless, it is worth noting that, with the exception of participant B, who noted frequent frustration with the LogKey keyboard and produced the least amount of data, and participant E, who was the oldest participant, that participants displayed dense patterning of their IKIs wherein each appeared to typically type at a speed within a range of 100 ms.

Participants additionally used a variety of applications during their participation, which confirmed self-reports from pre- and post-study interviews about the applications they use within a typical week. **Figure 3** below shows that while participants varied slightly in the degree to which they produced text through different applications, overall keybursts were primarily produced in messaging (55% of the total keybursts), followed by social media apps (21.8% of total keybursts). Furthermore, when examining the IKI of the first key in a keyburst produced in a given application (i.e., the inter-KB value), keybursts produced in messaging applications appeared to significantly predict lower inter-KB values when compared to keybursts in dating apps, email, note-taking apps, and browsers (see **Table 5**). Interestingly, messaging and social media applications did not appear to differ significantly from one another.

Of the 32,274 total keybursts produced by participants, 8,290 were unique keybursts, and 5,908 of those unique keybursts occurring only once. In other words, keybursts that occurred only once, such as *lolol*, make up a total of 18% of all keybursts produced in the study. The most frequently occurring keybursts were common stop words such *to*, *the*, *a*, and, the first- and second-person subject pronouns ranked first and fourth, respectively. Furthermore, the most frequently occurring variants of *Haha* and *Lol* (i.e., *haha* and *lol*) represented the 43rd and 61st most frequently occurring keybursts, respectively (See **Table 2**).

Frequency of *Haha* and *Lol*

As stated in the RQ1, this study sought to examine the frequency of occurrence of two paralinguistic variables, *Haha* and *Lol*. As previously mentioned, eight participants procured 908 individual transmitted messages containing 7,329 words for this study. This number includes 62 messages sent over WhatsApp or Google Chat by participant B and J, respectively (due to how they configured the texting settings on their phones), and excludes messages transmitted using MMS, Advanced Messaging, or messages that participants sent to third-parties who did not consent to participate in the study. Hand-coding of the 908 messages identified 32 tokens of *Haha* and 26 tokens of *Lol*, meaning that *Haha*'s relative frequency among all transmitted messages was 0.44% while *Lol*'s relative frequency was 0.35%. It is noteworthy that these relative frequencies are roughly similar to those found in Tagliamonte and Denis (2008) and Tagliamonte, (2016), although, as **Table 6** shows, the relative frequencies of *Haha* and *Lol* varied among individual participants. For example, *lol* made up 1.22% of all words participant J transmitted, and *haha* made up 0.79% of all words participant C transmitted. Further, three out of the six participants who transmitted either of these variables categorically transmitted one or the other.

As demonstrated in Meredith and Stokoe (2014), not all content composed for text-based messaging is sent. This study therefore sought to examine the frequencies of keybursts composed by participants as part of texting, regardless of whether those keybursts were transmitted or not. In total, 9,364 keybursts were composed by participants as part of constructing the 908 transmitted messages. This potentially means that, when comparing with the 7,329 words that were sent, that potentially 2,047 keybursts were composed *but deleted*

TABLE 5 | Linear Mixed Effects Models (with participant as random intercept) examining inter-KB and average intra-KB of keybursts across application type.

Predictors	**Inter -KB**		
	Estimates	**CI**	***p***
(Intercept)	200.72	164.59–236.84	<0.001
Browser (compared to messaging)	37.39	20.10–54.68	<0.001
Dating apps (compared to messaging)	–15.07	–20.98–-9.15	<0.001
Email (compared to messaging)	–30.07	–38.71–-21.44	<0.001
Note (compared to messaging)	14.29	6.81–21.77	<0.001
Social media (compared to messaging)	0.20	–3.14–3.54	0.906
Random effects			
σ^2		8,794.01	
τ_{00}		3,379.95 $_{\text{participant}}$	
ICC		0.2,776,377	
Observations		22,330	
Marginal R^2/Conditional R^2		0.004/0.280	

TABLE 6 | Frequencies of *Haha* and *Lol* by participants in sent texting data.

Participant	Haha (absolute/relative frequency)	Lol (absolute/relative frequency)	Total messages sent	Total words in messages
A	2/1.27%	1/0.63%	26	157
B	—	—	1	4
C	14/0.79%	10/0.56%	280	1769
D	—	—	30	416
F	2/0.16%	2/0.16%	104	1,234
H	14/0.62%	—	255	2,255
I	—	8/0.74%	151	1,086
J	—	5/1.22%	61	408
Total	**32/0.44%**	**26/0.35%**	**908**	**7,329**

TABLE 7 | Frequencies of *Haha* and *Lol* keybursts in texting data.

Participant	Haha (absolute/relative frequency)	Lol (absolute/relative frequency)	Total keybursts from Texting Data
A	2/0.97%	1/0.48%	207
B	—	—	4
C	17/0.75%	10/0.44%	2,281
D	1/0.23%	2/0.46%	430
F	3/0.26%	2/0.17%	1,154
H	19/0.57%	—	3,312
I	—	10/0.70%	1,437
J	—	6/1.11%	539
Total	**42/0.45%**	**31/0.33%**	**9,364**

by participants. A frequency matrix of the texting keybursts identified 3,009 unique keyburst as well as a counts of the absolute frequencies and calculations of the relative frequencies. The top 75 most frequent keybursts are shown in **Table 2**, which includes the variants *lol* and *haha* as the 60th and 74th most frequently occurring keybursts. Hand-coding of this matrix identified 18 variants of *Haha* and six variants of *Lol*, as shown in **Table 3**, which includes variants like *lolol* and *hahahahhh*. In total, this identified 42 variants of *Haha* (10 more than the transmitted tokens), or a relative frequency of 0.45% of texting keybursts, and 31 variants of *Lol* (5 more than the transmitted tokens), or a relative frequency of 0.33% of texting keybursts. This means that, while 10 tokens of *Haha* and five tokens of *Lol* were composed but never transmitted, their relative frequency is consistent across keyburst and transmitted data. Further, as shown in **Table 7**, participants varied with regard to how frequently they composed either variable, and, interestingly, this examination revealed that participant D indeed composed both *Haha* and *Lol* in spite of never sending either in a text message.

It is also important to note that both *Haha* and *Lol* were composed by participants in other applications besides texting. When examining the entire data set collected from LogKey, which includes keystrokes and keybursts from communicative and non-communicative applications, keybursts such as *lol* and *haha* maintain a similar level of relative frequency; occurrences of *lol* and *haha* keybursts (including their variants) therefore respectively represent 0.45% and 0.40% of all keybursts collected from this study.

TABLE 8 | *Haha* and *Lol* by turn-taking position.

Participant	Haha initial – terminal – medial			Lol initial – terminal – medial		
A		2		1		
C	9	4	1	3	5	2
F	1	1		2		
H	13		1			
I				1	6	1
J				3	2	
Total	**23**	**7**	**2**	**10**	**13**	**3**

Turn-Taking Structures

As discussed in ***Section 2.2***, paralinguistic features are commonly deployed at turn-taking positions in asynchronous messaging, which was verified by this study. Across participants, both *Haha* and *Lol* were predominantly used in the initial or terminal positions of a sent message. 23 tokens of *Haha* occurred in the initial position, seven in the terminal position, and two at a medial position between clause boundaries. 10 tokens of *Lol* occurred in the initial position, 13 in the terminal position, and three at a medial position between clause boundaries. While this study does not necessarily argue that *haha* and *lol* are variants of the same variable, a chi-square test of independence was performed to examine the relation between use of *haha* or *lol* at the initial or terminal position in a message. The relation between these was significant, X^2 (1, $N = 53$) = 6.1013, $p = 0.0134$, as *Haha* was more likely to be in the initial positions. As is shown in **Table 8**, individual participants varied in their use of these

TABLE 9 | Median intra-KB averages of *Haha* and *Lol* keybursts

Participant	Haha Tokens	Haha Median intra-KB Average (ms)	Lol Tokens	Lol Median intra-KB Average (ms)	Median intra-KB Average (ms) All Keybursts	Total Keybursts from Texting Data
A	2	336.0	1	74	103.75	207
C	17	212.2	10	248.7	159.2	2,281
D	1	62	2	67.97	59.25	430
F	3	210.4	2	233.5	84.33	1,154
H	19	71.00	—	—	52	3,312
I	—	—	10	123.8	114	1,437
J	—	—	6	196.7	176.3	539
Total	**42**	**135.38**	**31**	**176.6**	**85**	**9,364**

variables at the different positions. For example, Participant H nearly exclusively used *haha* in the initial position, and Participant I similarly tended to use *lol* in the terminal position, which suggests that the above chi-square test may have been biased by the observed habits of individual participants. The situational use of these features at these turn-taking positions will be further discussed in ***Section 5***.

Timing Processing of Composing *Haha* and *Lol*

As demonstrated in **Tables 4** and **5**, participants' articulatory timing processes varied, as would be expected from keystroke data. Indeed, while similar variations were found when examining participants' individual timing processes for composing *haha* and *lol* in texting data, in general participants tended to take more time to compose *haha* and *lol* than typical keybursts. As shown in **Table 9** below, across seven participants who composed *haha* or *lol* in the texting data, the median intra-KB average (i.e. the average speed of composing a keyburst) was 85ms, whereas *Haha* was 135.38 ms and *Lol* 176.6 ms.

That *Haha* and *Lol* keybursts appeared to have different timing patterns was confirmed when examining the inter-KB values (i.e., the time prior to composing the keyburst). A linear mixed effects model (see **Table 10**) showed that variants of *Haha* and *Lol* keybursts were more likely than all other keybursts to have higher inter-KB values (p = 0.056 and p = 0.017, respectively). A possible interpretation of this is that participants took more time to cognitively select these specific features and then more time to compose them because they were dedicating more cognitive attention to use of these symbolic features, especially when compared to all other keybursts.

This interpretation is further evidenced by the temporal data concerning *Haha* and *Lol* keybursts beyond texting data. As shown in **Figure 4**, participants composed *Haha* and *Lol* in texting and all other mobile applications during their participations, totaling 145 *Haha*s and 149 *lol*s across all keystroke data. However, welch two-sample t-tests indicated there was no significant difference between inter-KB values for either *Haha*s composed in messaging and non-messaging apps (t [83.224] = 0.16123, p = 0.8723) or *Lol*s in messaging and non-messaging apps (t [76.157] = 1.2898, p = 0.201). A possible interpretation of this is that these features are used similarly across other mobile writing registers, or that the other mobile apps that participants composed these features in bear resemblance to texting. Indeed, besides messaging apps, these features were composed in dating apps (4 *Haha*s, 22 *Lol*s) and social media apps (41 *Haha*s, 71 *Lol*s), which may share similar asynchronous messaging structures.

TABLE 10 | Linear Mixed Effects Model for inter-KB and Haha or Lol keybursts (participant as random intercept) in texting data (excluding keybursts with inter-KB > 2000ms).

	inter Total		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	266.04	216.67–315.41	**<0.001**
Laughter	56.50	−1.48–114.48	0.056
Lol	57.51	10.34–104.68	**0.017**
Random effects			
σ^2		66,222.53	
$\tau_{00\ \text{participant}}$		4,764.82	
ICC $_{\text{participant}}$		0.07	
Observations		8,687	
Marginal R^2/Conditional R^2		0.001/0.068	

Turn-Taking Structures and Timing

As discussed in ***Section 4.2.1***, participants overwhelmingly composed *Haha* and *Lol* keybursts in the initial or terminal positions of text-messages as opposed to medial positions within messages, and *Haha* was more likely to be used in the initial position. Interestingly, the timing of *Haha* and *Lol* appeared to differ at different turn-taking positions, as *Haha*s and *Lol*s in the initial position were generally composed faster than in the terminal position. As shown in **Table 11**, this is most dramatic for *Haha* keybursts, where the median intra-KB average was over 200 ms *slower* in the terminal position than the initial or medial positions. While in general any keyburst in the initial position was composed faster than the terminal position (103.5 ms vs. 125.5 ms), when excluding keybursts with inter-KB values above 2000 ms, a linear mixed effects model did not appear to show any significant relationship between message position and intra-KB average across all participants (p > 0.05). A possible interpretation of this is that the psychomotor processes for composing the first and last words in a text message may involve greater cognitive demands, especially when compared to

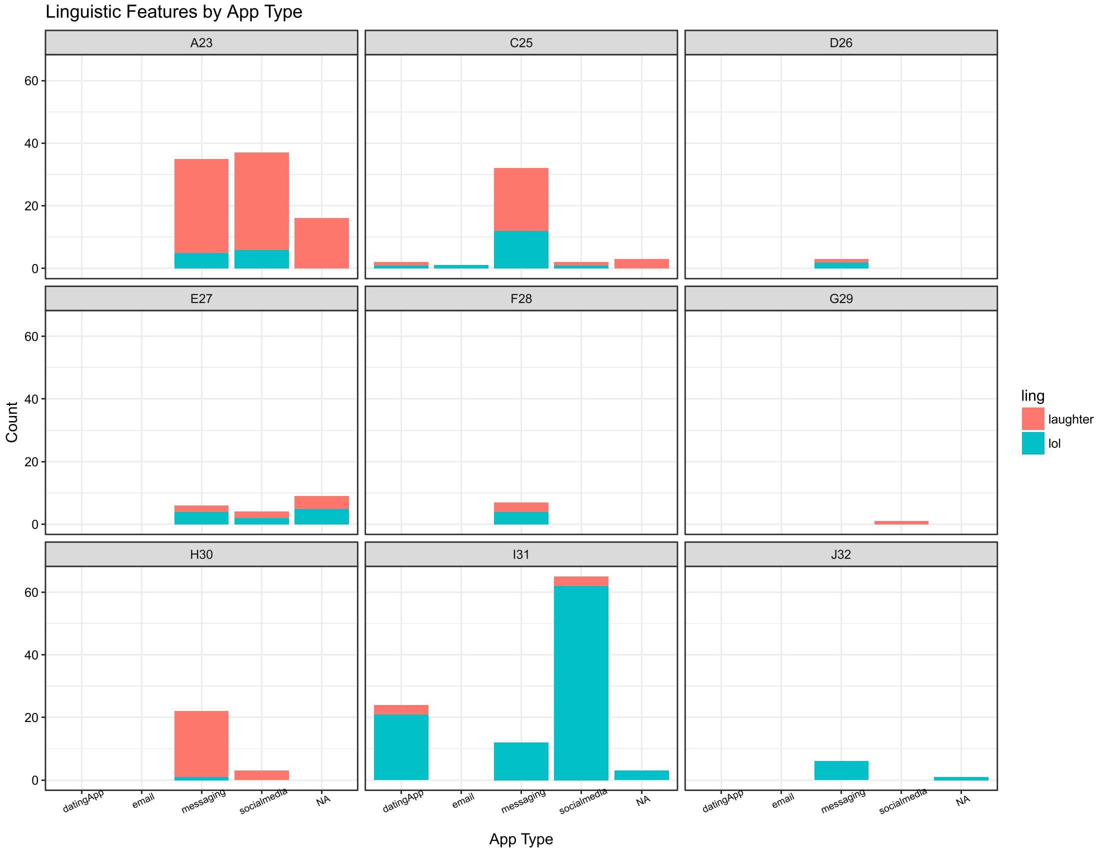


FIGURE 4 | Occurrence of Haha or Lol feature by application type for each participant.

TABLE 11 | Median intra-KB averages of *Haha* and *Lol* by turn-taking position.

Position	Haha Tokens	Haha Median intra-KB Average (ms)	Lol Tokens	Lol Median intra-KB Average (ms)	All Keybursts Median intra-KB Average (ms)
Initial	23	121.00	10	174.6	103.5
Terminal	7	328.5	13	185.0	125.5
Medial	2	107.98	3	135.67	80.25
Total	**32**	**143.9**	**26**	**174.6**	**85.0**

the general speed of composing medial keybursts, which would bear some similarity to speech rates toward the end of utterances (Sóskuthy and Hay, 2017).

In order to further investigate the relationship between composition of these paralinguistic features at different turn-taking positions, additional tokens were included from tokens of *Haha* and *Lol* composed while the social media app Snapchat. Snapchat is a social media app that involves both photo-sharing as well as instant-messaging (Ekman, 2015; Jeong and Lee, 2017; Ilbury, 2018). As previously mentioned, six participants used Snapchat during their participation in the study, producing a total of 41 *Haha*s and 71 *Lol*s. When examining the strings of keystroke and keyburst data from the contexts in which these tokens occurred, they indeed appeared to resemble discrete messages. While Snapchat messages disappear after viewing, making it impossible to know who the participants were communicating with or even what the transmitted message was, it was possible to identify 74 *likely* messages from keystroke data. This included 19 *Haha*s and 55 *Lol*s. As shown in **Table 12**, a linear mixed effects model suggests that, when aggregating both variables together across texting and Snapchat messages, *Haha* or *Lol* in the terminal position was a significant predictor of higher intra-KB averages. **Figure 5** visualizes this distinction as well. A possible explanation of this is that composing salient, paralinguistic features such as *haha* or *lol* at the end of a message on a mobile device, regardless

TABLE 12 | Linear mixed-effects model of Lol and Haha in Snapchat and SMS data (participant as random intercept).

	Average Intra-KB IKI		
Predictors	**Estimates**	**CI**	***p***
(Intercept)	214.59	130.35–298.83	**0.001**
Medial (vs. Initial)	−60.41	−144.66–23.85	0.162
Terminal (vs. Initial)	71.58	13.12–130.04	**0.018**
Texting (vs. Snap)	−25.98	−99.50–47.54	0.491
Random effects			
σ2		19,674.96	
τ00 participant		5,800.19	
ICC participant		0.23	
Observations		132	
Marginal R2/Conditional R2		0.086/0.294	

of the specific register, may involve different psychomotor processes related to how participants were cognitively attending to how that paralinguistic cue functioned in context to their ongoing discourse. As will be further discussed in the following section, I argue that Brown and Levinson's (1987) politeness framework offers one possible interpretation for this apparent distinction in digital articulatory processes.

ANALYSIS: KEYSTROKES IN CONTEXT

In the above ***Section 1*** provided an overview of the keystroke data collected from this study, which identified the frequencies and timing processes of the paralinguistic features *Haha* and *Lol* that participants composed while text-messaging, as well as how these features functioned as part of turn-taking structures in messaging. Importantly, this analysis found that *Haha* and *Lol* were predominantly composed at the start or end of a message, and that they were composed more quickly at the start of a message when compared to the end. In this Section I will demonstrate how, as Van Waes et al. (2015) suggested, keystroke analysis may be combined with qualitative discursive analysis in order to better understand the interactional contexts through which individuals engage in text-based mobile communication. This will involve examining the interactional contexts in which participants composed *Haha* or *Lol* through the lens of Brown and Levinson's (1987) politeness framework. In doing so, I hope to offer an approach that serves as one possible method to heed Nguyen et al.'s (2016) challenge to reconcile the analysis of macro-scale data channels obtained from computational methods (i.e., keystroke data) and discursive analysis of real-world social performances (i.e., sent text message data).

As discussed in ***Section 2.2***, according to Brown and Levinson (1987), individuals may use various politeness strategies during communication, such as *shared laughter* (Jefferson, 1979) in order to mitigate possible positive or negative FTAs to themselves and interlocutors. The paralinguistic features *Haha* or *Lol* may function similarly to *shared laughter* by being deployed at turn-taking positions in messaging as a means of managing politeness strategies within asynchronous text-based mobile registers like text-messaging. I further suggest that this frames keystroke data, which provides insights into temporal dimensions of digitally articulation, as reflecting attention-paid-to-text, or how individuals may strategically manage composed and sent message content to negotiate their relationships through texting. I will attempt to demonstrate this through example conversations from participant data, particularly from participants C and H.

Participant C, who produced 17 *Haha* tokens and 10 *Lol* tokens, routinely composed these features faster at the start of a message than at the end. When examining how C used many of these tokens

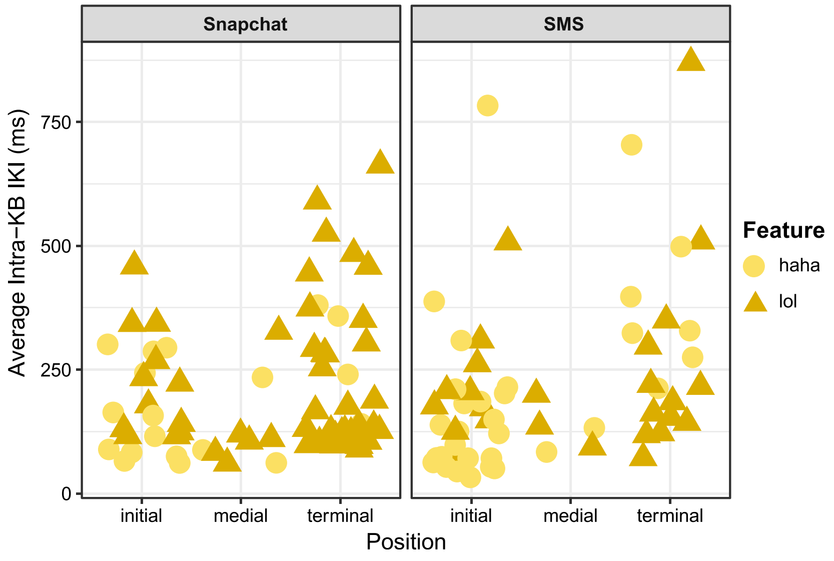


FIGURE 5 | Jitter plot showing average speed of Lol and Haha for SMS and Snapchat data.

in context, they offer vivid examples of how these features strategically management politeness, particularly when in conversation with their partner. In Example C01, C and their partner are discussing beverages to purchase for a social gathering they are planning. In turn 01, C's partner references a specific alcoholic beverage in what appears to be a joke (as indicated by *Lmao* and the hot face emoji at the initial and terminal turn-taking positions), and C's response with an initial *Haha*, "Haha oh lordy," in turn 02 therefore likely serves to indicate humorous reception and preserve their partner's positive face. This is further supported by the fact that the intra-KB average of this *Haha* was 181.5 ms.

In Example C02, while discussing the logistics of coordinating a future trip, C's partner declines C's suggestion in turn 02 ("That sounds like a process though lol"). In the next turn, C's partner offers a joke, perhaps to mitigate their FTA from the previous turn. C's response in turn 04 ("Its already a process"), which used some of their partner's language from turn 02, appears to reject the joke, yet still includes a *haha* in the terminal position, composed with an intra-KB average of 328.5 ms. This terminal *haha* therefore appears to mitigate the FTA of rejecting their partner's joke, and the slower composition time, when compared to the initial *Haha* in Example C01, may indicate that C deliberately attended to the content of their message in order to preserve their partner's positive face. In other words, the faster speed of composing the initial *Haha* may be indicative of C reacting to their partner's previous turn, but the slower speed of composing the terminal *haha* may indicate C was paying close attention to the message they were in the act of composing.

Across these as well as other examples, participant C appears to use initial and terminal paralinguistic features for different strategies as part of turn-taking structures. In another exchange with their partner, the terminal *haha* in the message "Shes important to my mom, but not so much me haha" was composed at an average intra-KB IKI of 498.25 ms. Similarly, the *lol* in "Lol nvm haha" was composed at 172.6 ms (and the *haha* in the terminal position was 212.25 ms), whereas in "Mom what lol" it was composed at 297.3 ms. A possible explanation for taking more time to compose these features in the terminal position may be that the composer of the message needs to review the content of their message and mitigate any positive face threats to the receiver. In other words, composing *Haha* at the initial position may be cognitively retrieved and physiologically inputted faster than at the terminal position *because* they serve alternative discursive functions. For example, if a *Haha* at the initial position may be managing the positive face of the interlocutor, perhaps in reaction to the interlocutor offering a joke or sacrificing their own face (as in the message "Haha oh lordy"), then the faster speed may be indicative of the compulsion to quickly react to and placate the positive face of the interlocutor. In other words, composing these speeds at different discursive positions may correspond with face management strategies and even how the sender is paying attention to how their message content. This would bear resemblance to shared laughter's function to coordinate the start of a responder's turn and their acceptance of appropriateness (Jefferson, 1979), as well as Brown and Levinson's (1987) observation that delayed responses in spoken turn-taking structures may signal a face threat based on prior turns.

Examining the keystroke data through the lens of politeness strategies also offers an explanation of revision patterns. As Meredith and Stokoe (2014) suggested, repairs in message construction, which are seen only by the individual composing the text, do more than correct textual errors—they may reflect how individuals are deliberately managing their interpersonal relationships. This may be demonstrated through participant H's individual compositional habits, who was the fastest writer among participants with a median intra-KB average speed of 52 ms. H generally used variants of *Haha* at or even below this value, and predominantly in the initial position. As show in Example H01, the initial *Haha* in turn 04 ("Haha all right so should i just wear workout clothes?") was composed at a rate of 50.5 ms, whereas in Example H02 the initial *Haha* in turn 04 ("Haha yea kinda i just have no idea whos coming or when but well see how it goes!!!") was written at a rate of 149.25 ms. H wrote each of these messages to the same interlocutor, identified as Friend11, and after closer examination of the content and context of their messages suggests that the nature of their relationship changed during H's week of participation in the study. For the first two days of H's participation, as shown in Example H01, H and Friend11 primarily coordinated social activities, what Thurlow and Brown (2003) termed the *social-arrangement* orientation. However, on the third day of participation, Friend11 appeared to have left on a family vacation, so the following messages, as shown in Example H02, concerned catching up on their daily activities at a distance, what Thurlow and Brown (2003) termed the *informational-relational* orientation. The change in the speeds at which H composed initial *Haha*s to Friend11 may reflect a shift in communication orientations, which required H to respond to turns from Friend11 with message content that would maintain their relationship. This orientation likely would have afforded H more time to respond to messages, since they weren't time sensitive in the way the social-arrangement orientation was, and is evidenced by the fact that more time elapses between turns in Example H02 than in H01.

Another explanation may be that H and Friend11's relational closeness was undergoing a shift. On the day that Friend11 appears to have left town, the two exchanged late-night messages in which Friend11 asked for H's Snapchat handle, as shown in Example H03. In their pre-study interview, H indicated that they typically message with their most intimate friends on Snapchat, so the addition of Snapchat as another channel for H and Friend11 to communicate through might signal such a potential shift in intimacy. This is further evidence by the message Friend11 sends in turn 06 ("But for real, this is random but you are amazing and im really glad we met. Spending time with you is so much fun"), and H's response in turn 07. Prior to exchanging closings salutations, the two then agree to send messages through Snapchat and text while Friend11 is out of town, as they do in Examples H02 and

Example C01

Turn	Interlocutor	Time sent	Message
01	Partner	12:43:04	Lmao well I already have my pear cider and pink lemonade [HOT FACE EMOJI]
02	C	12:43:19	Haha oh lordy
03	Partner	12:43:44	Oh it's going to be sensational

Example C02

Turn	Interlocutor	Time sent	Message
Turn	Interlocutor	Time sent	Message content
01	C	8:52:27	We can always do the week-of ticketd
02	Partner	8:53:05	That sounds like a process though lol
03	Partner	8:53:16	Like every week we have to see if we have to go to cleveland that week
04	C	8:53:25	Its already a process haha
05	Partner	8:54:30	This is true but at least we have the ability to do a little planning lol

Example C03

Turn	Interlocutor	Time sent	Message
Turn	Interlocutor	Time sent	Message
01	A	9:28:24	Guess who didnt get funding for next year
02	Partner	9:29:02	[Redacted]!
03	A	9:29:53	Me!
04	Partner	9:30:14	Yayjjj!!
05	Partner	9:30:16	Oh wait
06	A	9:30:26	Beb lol
07	Partner	9:30:32	Is ok babe. How are you feeling about it?

H04. In other words, the initial *Haha*s in Examples H02 and H04 may signal the potential change in their friendship. In composing these features more slowly (as well as writing more), it is possible that H was paying greater attention to the text they produced (Biber and Conrad, 2009; Kendall, 2013) in order to engage in more self-monitoring (Hayes, 2012; Bowen and Van Waes, 2020) to be strategic about what they write in sending or responding to messages, and this may have been done in order to maintain the possibility of further advancing their relationship with Friend11. Indeed, when more closely examining the sequence of keystrokes for how H composed the message in turn 02 in Example H04, H initially began composing the message *without* the *Hahah*. H *added* the *Hahah* as part of a message construction repair (Meredith and Stokoe, 2014), so the slower timing process for the *Haha* feature at the initial position (71.8ms) may also reflect greater cognitive effort as H strategically self-monitored their message content—including use of *Haha*—in order to preserve their changing relationship with Friend11.

This closer examination of how participants C and H composed *Haha* or *Lol* at different positions at different speeds offers a window into the complex and variable ways that these individuals may meaningfully manage their relationships through the psychomotor process of inputting text. In particular, this examination suggests that the timing processes of these features may indeed reflect the discursive conditions under which individuals are working, through the socio-technical and even psychomotor mechanisms, to manage facework; and this may correspond with the structural positioning of paralinguistic features that are intended to symbolically manage face. Further, this finding simultaneously demonstrates the value in combining closer qualitative analysis amid the breadth of data obtained from keystroke-logging methods.

DISCUSSION

This study used keystroke-logging methods in order to examine how paralinguistic variables were composed as part of text-based, mobile communication. Its research questions asked, based on keystroke data, 1) What are the frequencies of occurrence of paralinguistic variables *Haha* and *Lol* in text messaging, and what is their distribution according to turn-taking structures? And, 2) What are the timing processes of *Haha* and *Lol* in text messaging, and how do these timing processes reflect turn-taking structures in asynchronous messaging? This study found that, for the participants that sent either of these features in texting (N = 6), pauses before composing *Haha* and *Lol* were significantly greater

Example H01

Turn	Interlocutor	Time sent	Message
01	Friend11	16:39:51	Want to go out at like 2–3?
02	H	16:40:39	Yeah lets do that
03	H	16:41:52	Also again ive never been there for the rafting or hiking or courses so i have no preference or idea what we should do
04	Friend11	17:55:55	Haha ill show you around its all good
05	H	17:57:51	Haha all right so should i just wear workout clothes?
06	Friend11	18:07:43	Yeah that sounds great. About to play our gig so ill text you after. Hope youve had a good day)
07	H	18:11:31	Good luck!!! yeah let me know how it goes!

Example H02

Turn	Interlocutor	Time sent	Message
Turn	Interlocutor	Time sent	Message
01	Friend11	12:30:33	How are you?
02	H	16:10:59	Good!! todays gone by so fast
03	Friend11	16:52:34	Yeah todays been a good one, we went to the montreal botanical gardens which sounds boring but it was actually pretty sweet. Getting excited for tonight?
04	H	16:58:48	Haha yea kinda i just have no idea whos coming or when but well see how it goes!!!
05	Friend11	17:06:45	Wish I could be there, hope its alot of fun! what did you and rebecca do today?
06	H	18:06:58	We drove around and then sat at some coffee place and did work then she went to some yoga place and now shes back so were about to start getting things ready
07	Friend11	19:03:31	Gotcha sounds like it was a great day. At dinner ill text you after hope its #lit

than for all other kinds of keybursts (see **Table 6**). Since these features also occurred with consistent, low frequency at initial and terminal turn-taking positions in a message, I suggest a possible interpretation of this finding is that *Haha* and *Lol* are established and discursively salient features for these participants, and the longer pause prior to composing these features suggests these participants are expounding more cognitive resources to carefully attend to how they use these features.

This study also found that, when comparing the speeds at which *Haha* and *Lol* were produced at the starts and ends of messages in SMS and Snapchat ($N = 132$), that these keybursts were produced faster at the start of a message than at the end of the message. A possible explanation for this may be that the different timing processes is reflective of the use of these paralinguistic features to signal different politeness strategies at the start and end of a message. In other words, composing *Haha* or *Lol* at the start of a message may be faster because participants were reacting to the message from their interlocutor's previous turn, whereas composing these features at the end of a message may be slower because participants were self-monitoring their own message to maintain their own positive face.

These findings support theories of cognitive processes of writing, particularly for the processes of producing and monitoring textual production (Hayes, 2012; Leijten et al., 2014), as well as Kendall's (2013) and Biber and Conrad's (2009) suggestions that greater attention to producing spoken or written language may affect the timing of those articulatory processes. Further, while this study was concerned with paralinguistic features typically represented in discrete word units, these findings do suggest that, as Eisenstein (2015) posits, that the production of text reflects the residue of spoken language production. As demonstrated by these methods for conducting linguistic analysis from keystroke data, it may therefore prove fruitful to examine morpho-syntactic or morpho-phonological variables to examine the extent to which, as Eisenstein (2015) speculates, this residue permeates to different linguistic levels.

Limitations

While I maintain that the findings of this study remain promising, I in no way argue these findings are generalizable due to its limitations. Indeed, this was the first study to use LogKey, and one of the first—if not the first—to study keystroke-logging analysis from individuals' personal mobile devices for the period of one week. This study therefore simultaneously functioned as a test of its methodological feasibility, or, to paraphrase the software engineer to helped design LogKey, this study was like a beta test. This study was therefore methodologically limited in two ways: first, in terms the design of LogKey, and in the amount of data yielded from this study. LogKey's design and logging methods could therefore be improved to be more usable, and to log additional records. For example, logging when an application is opened on a device, when SMS messages are transmitted, and when the virtual keyboard appears on a mobile device's screen would all aid in more precisely determining when users might read a received message, and begin to plan composing a response to that message. Further, log records that include information regarding approximate GPS coordinates or gyroscope sensors would provide information regarding whether participants may be in private or public spaces as well as how they might be moving about those spaces.

Example H03

Turn	Interlouctor	Time sent	Turn
01	Friend11	1:21:12	Whats your snap
02	H	1:21:33	Hahahahhh
03	H	1:21:38	[Redacted]
04	Friend11	1:21:46	U Got
05	Friend11	1:21:49	Ok
06	Friend11	1:23:19	But for real, this is random but you are amazing and im really glad we met. Spending time with you is so much fun
07	H	1:27:26	No I know but it was so random how it started and I am too it's gonna suck not being able to hang out for the next week
08	H	1:27:31	I agree
09	Friend11	1:28:25	I meant like me saying that was random haha
10	H	1:28:46	No no I know
11	H	1:29:00	But then I decided to include that little detail
12	Friend11	1:29:01	But yeah im going to miss you. Ill snap or text you when I get to [REDACTED]
13	Friend11	1:29:07	Oh beastttt
14	H	1:29:17	Thank u for that
15	Friend11	1:29:31	Youre very welcome
16	H	1:29:44	And yea snap me

Example H04

Turn	Interlouctor	Time sent	Turn
Turn	Interlocutor	Time sent	Message
01	Friend11	19:03:31	How was last night?
02	H	10:49:41	Hahah i mean it was good and fun but i was kinda bored and wanted people to just leave
03	Friend11	10:56:53	Hahaha classicc
04	Friend11	11:04:51	Headed to the train station to go to [REDACTED]

Future Directions

In addition to addressing the limitations discussed above, future study could continue to examine the data collected in different ways. For example, analysis of revision patterns could examine whether patterns found by Bowen and Van Waes (2020) may extend to texting, analysis of the timing processes of morphophonological features, such as -t/-d or -ing deletion, may examine Einstein's (2015) observations from Twitter data, and more detailed analysis of *pause bursts* may yield greater insight into how the psychomotor processes of composing on mobile devices may differ from writing on computer-based keyboards (Mangen et al., 2015; Galbraith and Baaijen, 2019). Such analyses would necessarily expand upon the tools used for this study, requiring means to parse keystroke data in order to identify revision bursts and morphological features from keybursts.

At the same time, future study could apply these methods of combining linguistic analysis with keystroke data to examine other mobile writing registers. For example, writing on Twitter, Snapchat, Instagram, or other forms of social media would be valuable not only to examine the writing processes on these registers, but also perhaps also allow examination of emerging text-based linguistic features. Further, as was evident from participants in this study, while text-messaging is predominantly text-based, use of emojis and other visual elements is increasingly valuable, therefore developing means to examine the production of text in conjunction with use of visual symbolic material (see Ilbury, 2018) may paint a different picture than examining text alone. Of course, all of these possible future directions would require not only a more robust version of LogKey, or a similar keystroke logger, but also coordination with other data channels. For example, as Olive and Cislaru (2015) demonstrated, combining corpus-based methods for collected large data sets from Twitter and keystroke-logging methods may offer valuable understanding into how text-based linguistic features are diffused *and* recontextualized by individuals. Such a study would have the ability to more powerfully understand how individuals perform the individual through the social.

Coda

I hope I have demonstrated the potential value of computational methods such as keystroke logging, especially for examining text-based data. As I have laid out above, the development of keystroke analysis methods for examining text-based data produced through mobile technology offers one possible route to conduct sociolinguistic research in the 21st century. Through such a methodology, linguistic units not just asynchronous artifacts, but may be again seen as evidence of flesh-and-blood processes for articulating language in real-time.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available per the privacy, data protection policy, and informed consent process that was approved by the North Carolina State

University's Institutional Review Board, data from this study may not be shared.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by North Carolina State University Institutional Review Board (Protocol 11651). The participants provided their written informed consent to participate in this study.

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Ælfricof, A. E., and Zupitza, J. (1880). Ælfric's grammatik und glossar (J. Zupitza, Trans.). Weidmannsche Buchhandlung Retrieved from: https://archive.org/details/grammatik00aelfuoft/page/n287/mode/2up.

Andre, B. K., Ording, B., and Christie, G. (2005). Activating virtual keys of a touch-screen virtual keyboard. Cupertino, CA, USA: Apple Inc., 7,844.914.B2.

Androutsopoulos, J. (2014a). Languaging when contexts collapse: audience design in social networking. *Discourse, Context & Med*. 4–5, 62–73. doi:10.1016/j.dcm.2014.08.006

Androutsopoulos, J. (2014b). "Mediatization and sociolinguistic change: key concepts, research traditions, open issues," in *Mediatization and sociolinguistic change*. Editor J. Androutsopoulos (Boston, MA: De Gruyter), 3–48.

Androutsopoulos, J. (2014c). Moments of sharing: entextualization and linguistic repertoires in social networking. *J. Pragmat*. 73 (1), 4–18. doi:10.1016/j.pragma.2014.07.013

Baron, N. S., and Ling, R. (2011). Necessary smileys & useless periods. *Visible Lang*. 45 (1–2), 45–67.

Baron, N. S. (2004). See you online: gender issues in college student use of instant messaging. *J. Lang. Soc. Psychol*. 23 (4), 397–423.

Baym, N. K., Zhang, Y. B., and Lin, M. (2004). Social interactions across media: interpersonal communication on the internet, telephone, and face-to-face. 6. *New Media Soc*. (3), 299–318. doi:10.1177/1461444804041438

Biber, D., and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Bourdieu, P. (1984). *Distinction: a social critique of the judgment of taste*. Abingdon, United Kingdom: Routledge.

Bowen, N., and Van Waes, L. (2020). Exploring revisions in academic text: closing the gap between process and product approaches in digital writing. *Writ. Commun*. 37 (3), 322–364. doi:10.1177/0741088320916508

Brock, A. (2020). *Distributed blackness: african American cybercultures*. New York, NY: NYU Press.

Brown, P., and Levinson, S. C. (1987). *Politeness: some universals in language usage*. Cambridge, United Kingdom: Cambridge University Press.

Butler, J. (1997). *Excitable speech: a politics of the performative*. New York, NY: Routledge.

Cislaru, G. (2015). *Writing(s) at the crossroads* (Philadelphia, PA: John Benjamins Publishing Company).

Coupland, N. (2001). "Language, situation, and the relational self: theorizing dialect-style in sociolinguistics," in *Style and sociolinguistic variation*. Editors P. Eckert and J. R. Rickford (Cambridge, United Kingdom: Cambridge University Press), 185–210.

Coupland, N. (2007). *Style language variation and identity*. Cambridge, MA: Cambridge University Press.

Eckert, P. (2000). *Linguistic variation as social practice: the linguistic construction of identity in Belten High*. Malden, MA: Wiley-Blackwell.

Eckert, P. (2001). "Style and social meaning," in *Style and sociolinguistic variation*. Editors P. Eckert and J. R. Rickford (Cambridge, United Kingdom: Cambridge University Press), 119–126.

Eckert, P. (2008). Variation and the indexical field. *J. SocioLinguistics* 12 (4), 453–476. doi:10.1111/j.1467-9841.2008.00374.x

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS One* 9 (11), e113114. doi:10.1371/journal.pone.0113114

Eisenstein, J. (2013). Phonological factors in social media writing. in Naacl 2013: proceedings of the workshop on language analysis in social media, Atlanta, Georgia, June 13, 2013, 11–19. Stroudsburg, Pa: Association for Computational Linguistics. Available at: http://www.cc.gatech.edu/~jeisenst/papers/lasm13-phono.pdf.

Eisenstein, J. (2015). Systematic patterning in phonologically- motivated orthographic variation. *J. SocioLinguistics* 19 (2), 161–188. doi:10.1111/josl.12119

Ekman, U. (2015). Complexity of the ephemeral - snap video chats. *Empedocles* 5 (1–2), 91–101. doi:10.1386/ejpc.5.1-2.97_1

Farman, J. (2012). *Mobile interface theory: embodied space and locative media*. New York, NY: Routledge.

Fuchs, S., Savin, E., Solt, S., Ebert, C., and Krifka, M. (2019). Antonym adjective pairs and prosodic iconicity: evidence from letter replications in an English blogger corpus. *Linguistics Vanguard*. 5 (1). doi:10.1515/lingvan-2018-0017

Galbraith, D., and Baaijen, V. (2019). "Aligning keystrokes with cognitive processes in writing," in *Observing writing: insights from keystroke logging and handwriting*. Editors E. Lindgren and Sullivan Kirk.. (Leiden, Netherlands: Brill), 306–325. doi:10.1163/9789004392526_015

Georgakopoulou, A. (2014). "Girlpower or girl (in) trouble? identities and discourses in the (new) media engagements of adolescents' school-based interaction," in *Mediatization and sociolinguistic change*. Editor J. Androutsopoulos (Berlin, Boston: De Gruyter), 217–244.

Goffman, E. (1956). *The presentation of self in everyday life*. Edinburgh, Scotland: University of Edinburgh Social Science Research Centre.

Grésillon, A., and Perrin, D. (2015). "Methodology: investigating real-life writing processes," in *Writing(s) at the crossroads*. Editor G. Cislaru (Philadelphia, PA: John Benjamins Publishing Company), 33–54.

Grieve, J., Nini, A., and Guo, D. (2017). Analyzing lexical emergence in modern American English online. *Engl. Lang. Ling*. 21 (1), 99–127. doi:10.1017/s1360674316000113

Haas, C., Takayoshi, P., Carr, B., Hudson, C., and Pollock, R. (2011). Young people's everyday literacies: the language features of instant messaging. *Res. Teach. Engl*. 45 (4), 378–404.

Hayes, J. R. (1996). "A new framework for understanding cognition and affect in writing," in *The science of writing: theories, methods, individual differences, and applications*. Editors C. M. Levy and S. Ransdell (Newark, Delaware: International Reading Association), 1–27.

Hayes, J. R. (2012). Modeling and remodeling writing. *Writ. Commun*. 29 (3), 369–388. doi:10.1177/0741088312451260

Highfield, T., and Leaver, T. (2016). Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Commun. Res. Practice* 2 (1), 47–62. doi:10.1080/22041451.2016.1155332

Ilbury, C. (2018). *Self-presentation and the sociolinguistics of Snapchat*. Paper presented at BAAL Language & New Media SIG 2018, England, United Kingdom: The Open University.

Jefferson, G. (1979). "A technique for inviting laughter and its subsequent acceptance declination," in *Everyday language: studies in ethnomethodology*. Editor G. Psathas (New York, NY: Irvington), 79–95.

Jeong, D. C., and Lee, J. (2017). Snap back to reality: the cognitive mechanisms underlying Snapchat. *Comput. Hum. Behav*. 77, 274–281. doi:10.1016/j.chb.2017.09.008

Jones, T. (2015). Toward a description of African American vernacular English dialect regions using "Black Twitter. *Am. Speech*. 90 (4), 403–440. doi:10.1215/00031283-3442117

Jurgens, D. (2013). That's what friends are for: inferring location in online social media platforms based on social relationships. in Proceedings of the seventh international AAAI conference on weblogs and social media, Cambridge, MA, July 8–11, 2013, Association for the Advancement of Artificial Intelligence. 273–282.

Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: studies in corpus sociophonetics*. London, United Kingdom: Palgrave Macmillan.

Laursen, D. (2005). "Please reply! the replying norm in adolescent SMS communication," in *The inside text: social, cultural and design perspectives on SMS*. Editors R. Harper, L. Palen, and A. Taylor (Norwell, MA: Springer), 53–73.

Ledbetter, A. M., and Mazer, J. P. (2014). Do online communication attitudes mitigate the association between Facebook use and relational interdependence? An extension of media multiplexity theory. *New Med. Soc*. 16, 806–822. doi:10.1177/1461444813495159

Ledbetter, A. M. (2008). Media use and relational closeness in long-term friendships: interpreting patterns of multimodality. *New Media Soc*. 10 (4), 547–564. doi:10.1177/1461444808091224

Leijten, M., Macken, L., Hoste, V., Van Horenbeeck, E., and Van Waes, L. (2012). "From character to word level: enabling the linguistic analyses of Inputlog process data." in European association for computational linguistics, EACL – computational linguistics and writing (CL&W 2012): linguistic and cognitive aspects of document creation and document engineering, Avignon, France, April 24, 2012, Editors M. piotrowski, C. mahlow, and R. Dale. Available at: http://aclweb.org/anthology/W/W12/W 12-03.pdf.

Leijten, M., Van Horenbeeck, E., and Van Waes, L. (2019). "Analyzing keystroke logging data from a linguistic perspective," in *Observing writing: insights from keystroke logging and handwriting*. Editors E. Lingren and Sullivan K. P.H. (Boston, MA: Brill), 71–95.

Leijten, M., and Van Waes, L. (2013). Keystroke logging in writing research: using Inputlog to analyze and visualize writing processes. *Writ. Commun*. 30 (3), 358–392. doi:10.1177/0741088313491692

Leijten, M., Van Waes, L., Schriver, K., and Hayes, J. R. (2014). Writing in the workplace: constructing documents using multiple digital sources. *J. Writ. Res*. 5 (3), 285. doi:10.17239/jowr-2014.05.03.3

Leijten, M., Van Waes, L., and Van Horenbeeck, E. (2015). "Analyzing writing process data: a linguistic perspective," in *Writing(s) at the crossroads*. Editor G. Cislaru (Philadelphia, PA: John Benjamins Publishing Company), 277–302.

Lenhart, A. (2015). *Teens, social media & technology overview 2015: smartphones facilitate shifts in communication landscape for teens*. Retrieved from Pew Research Center website: http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/.

Lewis, C., and Fabos, B. (2005). Instant messaging, literacies, and social identities. *Read. Res. Q*. 40, 470–501. doi:10.1598/rrq.40.4.5

Ling, R. (2008). *New tech, new ties: how mobile communication is reshaping social cohesion*. Cambridge, MA: MIT.

Ling, R., and Yttri, B. (2001). "Hyper-coordination via mobile phones in Norway," in *Perpetual contact: mobile communication, private talk, public performance*. Editors J. E. Katz and M. A. Aakhaus (Cambridge, United Kingdom: Cambridge University Press), 139–169.

Mangen, A., Anda, L. G., Oxborough, G. H., and Brønnick, K. (2015). Handwriting versus keyboard writing: effect on word recall. *J. Writ. Res*. 7 (2), 227–247. doi:10.17239/jowr-2015.07.02.1

McCulloch, G. (2019). *Because internet: understanding the new rules of language*. New York, NY: Riverhead.

Meredith, J., and Stokoe, E. (2014). Repair: comparing Facebook "chat" with spoken interaction. *Discourse Commun*. 8 (2), 181–207. doi:10.1177/1750481313510815

Miller, K. S. (2006). "The pausological study of written language production," in *Computer keystroke logging and writing: methods and applications*. Editors K. P. H. Sullivan and E. Lindgren (Amsterdam, Netherlands: Elsevier), 11–30.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: a survey. *Comput. Ling*. 42 (3), 537–593. doi:10.1162/COLI_a_00258

Nguyen, D., Trieschnigg, D., and Cornips, L. (2015). Audience and the use of minority languages on twitter. in Proceedings of the AAAI international conference on web and social media, Boston, MA, May 26–29, 2015, program chairs M. Cha, C. Mascolo, and C. Sandvig. Menlo Park, CA: Association for the Advancement of Artificial Intelligence. Available at: http://www.dongnguyen.nl/publications/nguyen-icwsm2015.pdf.

Nottbusch, G. (2010). Grammatical planning, execution, and control in written sentence production. *Read. Writ*. 23 (7), 777–801. doi:10.1007/s11145-009-9188-4

Olive, T., and Cislaru, G. (2015). "Linguistic forms at the process-product interface: analyzing the linguistic content of bursts of production," in *Writing(s) at the crossroads*. Editor G. Cislaru (Philadelphia, PA: John Benjamins Publishing Company), 99–123.

Parisi, D. (2018). Archeologies of touch: interfacing with haptics from electricity to computing. Minnesota Press.

Pavalanathan, U., and Eisenstein, J. (2015a). Audience-modulated variation in online social media. *Am. Speech*. 90 (2), 187–213. doi:10.1215/00031283-3130324

Pavalanathan, U., and Eisenstein, J. (2015b). "Confounds and consequences in geotagged Twitter data," in *Empirical methods in natural language processing. Proceedings of the 2015 conference on empirical methods in natural language processing*. Editors L. Màrquez, C. Callison-Burch, and J. Su (Lisbon, NY: Association for Computational Linguistics), 2138–2148.

Pérez-Sabater, C. (2019). "Emoticons in relational writing practices on WhatsApp: some reflections on gender," in *Analyzing digital discourse*. Editors P. Bou-Franch and P. Garcés-Conejos Blitvich (Cham, Switzerland: Palgrave Macmillan), 163–189.

Piwek, L., and Joinson, A. (2015). "What do they snapchat about?" Patterns of use in time-limited instant messaging service. *Comput. Hum. Behav*. 54, 358–367. doi:10.1016/j.chb.2015.08.026

Plane, S. (2015). "Some problems encountered in the description and analysis of the dynamics of writing," in *Writing(s) at the crossroads*. Editor G. Cislaru (Philadelphia, PA: John Benjamins Publishing Company), 21–32.

Plank, B. (2016). "Keystroke dynamics as signal for shallow syntactic parsing." in Proceedings of the 25th international conference on computational linguistics (COLING). Osaka, Japan, December 11, 2016. 609–619.

Sampietro, A. (2016). Emoticonos y multimodalidad. El uso del pulgar hacia arriba en WhatsApp. *Aposta: Rev. Cien. Soc*. 69, 271–295.

Schneier, J., and Kudenov, P. (2018). Texting in motion: keystroke logging and observing synchronous mobile discourse. *Mobile Media Commun*. 6 (3), 309–330. doi:10.1016/j.chb.2015.08.026

SMeiTi (2017). *SMS to text*. Hong Kong, China.

Sóskuthy, M., and Hay, J. (2017). Changing word usage predicts changing durations in New Zealand English. *Cognition* 166, 298–313.

Spilioti, T. (2011). "Beyond genre: closings and relational work in texting," in *Digital discourse: language in the new media*. Editors C. Thurlow and K. Mroczek (Oxford: Oxford University Press), 67–85.

Suchman, L. A. (1987). *Plans and situated actions: the problem of human-machine communication*. New York, NY: Cambridge University Press.

Tagg, C. (2012). *Discourse of text messaging*. Bloomsbury UK: GB.

Tagg, C., Baron, A., and Rayson, P. (2012). I didn't spell that wrong did i. oops": analysis and normalization of SMS spelling variation. *Lingvisticoe Investigatione* 35 (2), 367–388. doi:10.1075/bct.61.12tag

Tagliamonte, S. (2016). So sick or so cool? The language of youth on the internet. *Lang. Soc*. 45, (1) 1–32. doi:10.1017/S0047404515000780

Tagliamonte, S. A., and Denis, D. (2008). Linguistic ruin? lol! instant messaging and teen language, in *Am. Speech* 83 (1), 3–34.

Thurlow, C., and Brown (2003). Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Anal. Online* 1 (1), 1.

Trudgill, P. (1974). *Sociolinguistics: an introduction*. Baltimore, MD: Penguin.

Van Waes, L., Leijten, M., Lindgren, E., and Wengelin, Å. (2015). "Keystroke logging in writing research: analyzing online writing processes," in *Handbook of writing research*. Editors C. A. MacArthur, S. Graham, and J. Fitzgerald. 2nd ed. (New York, NY: Guilford Press), 410–426.

Van Waes, L., Leijten, M., and Quinlan, T. (2010). Reading during sentence composing and error correction: a multilevel analysis of the influences of task complexity. *Read. Writ*. 23, 803–834. doi:10.1007/sm45-009-9190-x

Van Waes, L., Leijten, M., van Horenbeeck, E., and Pauwaert, T. (2012). *A generic XMLstructure for logging human computer interaction*. Paper presented at the 13th International EARLI SIG Writing Conference. Porto, Portugal, Jun 29, 2013.

Westerman, W. C., and Elias, J. G. (2006). Capacitive sensing arrangement. in *Apple Inc* (USA), 7,764.274B2.

WSJTech (2014). "*Mobile vs. desktop: 85% of time Twitter users spent on Twitter happened on a mobile device*. http://on.wsj.com/1gSsENd." [Twitter post]. Retrieved from: https://twitter.com/wsjtech/status/451886622788055040?lang=en.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 26 January 2021
doi: 10.3389/frai.2020.547531

That's Cool. Computational Sociolinguistic Methods for Investigating Individual Lexico-grammatical Variation

Hans-Jörg Schmid[1]*, *Quirin Würschinger*[1], *Sebastian Fischer*[2] *and Helmut Küchenhoff*[2]

[1]*Department of English and American Studies, LMU, Munich, Germany,* [2]*Statistical Consulting Unit StaBLab, Department of Statistics, LMU, Munich, Germany*

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Lala Hajibayova,
Kent State University, United States
Andrea Nini,
The University of Manchester,
United Kingdom
Florent Perek,
University of Birmingham,
United Kingdom

***Correspondence:**
Hans-Jörg Schmid
hans-joerg.schmid@lmu.de

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *31 March 2020*
Accepted: *29 September 2020*
Published: *26 January 2021*

Citation:
Schmid H-J, Würschinger Q, Fischer S and Küchenhoff H (2021) That's Cool. *Computational Sociolinguistic Methods for Investigating Individual Lexico-grammatical Variation. Front. Artif. Intell. 3:547531. doi: 10.3389/frai.2020.547531*

The present study deals with variation in the use of lexico-grammatical patterns and emphasizes the need to embrace individual variation. Targeting the pattern THAT'S ADJ (as in *that's right*, *that's nice* or *that's okay*) as a case study, we use a tailor-made Python script to systematically retrieve grammatical and semantic information about all instances of this construction in BNC2014 as well as sociolinguistic information enabling us to study social and individual lexico-grammatical variation among speakers who have used this pattern. The dataset amounts to 4,394 tokens produced by 445 speakers using 159 adjective types in 931 conversations. Using detailed descriptive statistics and mixed-effects regression models, we show that while the choice of some adjectives is partly determined by social variables, situational and especially individual variation is rampant overall. Adopting a cognitive-linguistic perspective and relying on the notion of entrenchment, we interpret these findings as reflecting individual speakers' routines. We argue that computational sociolinguistics is in an ideal position to contribute to the data-driven investigation of individual lexico-grammatical variation and encourage computational sociolinguists to grab this opportunity. For the routines of individual speakers ultimately both underlie and compromise systematic social variation and trigger and steer well-known types of language change including grammaticalization, pragmaticalization and change by invited inference.

Keywords: individual variation, lexico-grammatical variation, social variation, corpus data, mixed-effects regression models, language change word count: 10,380

INTRODUCTION

Sociolinguistics, both "traditional" and computational, has focused on regionally, socially and situationally conditioned variation on the linguistic levels of phonology and morphosyntax. Deviating from this tradition, we investigate individual variation on the interface between lexis and grammar. Our main goal is to demonstrate that having a closer look at individual variation–rather than treating it as noise or residual variance–can contribute to a better understanding not only of regional and social variation but also of lexical, pragmatic and grammatical variation and language change.

Of course we are not the first to take a keen interest in individual variation in the use of linguistic features and patterns. In forensic linguistics and author identification studies (Coulthard 2004), individual differences regarding the use and frequency of linguistic patterns have taken center stage for some time. Milestone publications highlighting individual differences in the field of sociolinguistics include Guy (1980); Wolfram and Beckett (2000); Tagliamonte and Baayen (2012) and Walker and

Meyerhoff (2013). However, the survey given by Walker and Meyerhoff (2013) shows two things: first, while many studies in variationist sociolinguistics, in fact starting as early as with Labov (1966), have acknowledged the importance of individual variation in principle, none of them have actually investigated the nature of individual variation and its implications in detail. And second, lexical or lexico-grammatical variation has not been addressed so far.

Individual differences regarding the mental representation of linguistic knowledge are the main concern of studies in the field of usage-based cognitive linguistics, for example by Barlow (2013) and Verhagen et al. (2018). In a similar vein, Dąbrowska (2015); Dabrowska (2016) has focused on individual differences in the language attainment of native and L2 speakers. Since individual speakers are the ultimate carriers of language change, it is not surprising that individual variation has been gaining increasing attention in corpus-based diachronic linguistics. Relevant publications include Gries and Hilpert (2010); Schmid and Mantlik (2015); Baxter and Croft (2016); Petré and Van de Velde (2018); Anthonissen (2020a); Anthonissen (2020b); Petré and Anthonissen (2020). Work in this tradition tends to be based on the assumption that frequency distributions in the works of individual authors can, under certain circumstances, be interpreted with regard to the writers' underlying mental representations.

Taking insights from these fields into consideration, the present paper aims to encourage researchers in computational sociolinguistics to embrace the study of individual differences in lexico-grammatical variation. It is not our main intention to provide an in-depth investigation of the pattern under investigation, i.e., the pattern *THAT'S ADJ*. Instead, we are using the pattern as an example to showcase potential methods to be extended in future work in computational sociolinguistics and to emphasize the relevance of studies of this type for understanding linguistic variation and change.

The paper is structured as follows. In *The Target Pattern: THAT'S ADJ* we will describe the lexico-grammatical pattern chosen to serve as a target of the present case study, the pattern THAT'S ADJ as in *that's right* or *that's nice*. *Data* will report the computational methods developed to retrieve the kind of data required for the study of individual lexico-grammatical variation. *Results: descriptive data summary* will provide a descriptive statistical summary of the results regarding social, situational and individual variation. *Inferential statistics and results* will present the inferential-statistical techniques we have used to gauge the influence of social, situational and individual factors on the use of the pattern. *Discussion* will discuss the cognitive implications of our results and the role of individual variation vis-à-vis social variation and various types of language change.

THE TARGET PATTERN: THAT'S ADJ

The pattern investigated in this article is illustrated in examples (1) to (6), taken from BNC2014. Each example is related to one of six dominant usage types of the pattern.

(1) 'Evaluative' use:
S0255: er you just everything's taken off you ev- totally everything's taken er ou- off your hands.
S0315: that's fantastic (S28F)

(2) 'Epistemic' use:
S0519: in-interestingly the we are having the last two summers certainly worse summer because the Gulf Stream has shifted south.
S0520: mm
S0521: **that's true** yeah (S24E)

(3) 'Ethical' use:
S0337: outside --ANONnameM's grandad's house you know there's always cars there (.) someone was in like a Ford Focus and like maybe a Ford Fiesta and like er she clearly did n't know how big her car was it was like full on not going anywhere and er would n't go past a parked car.
S0336: **that's mean** (S985)

(4) 'Emotive' use:
S0585: yeah for er yeah exactly yeah and I was like ugh that is so horrible and she's like yeah I threw up through my nose and I was like no
S0587: that's horrible (SNXG)

(5) 'Descriptive' use:
S0179: yeah (.) yeah (.) my opinion of him went down.
S0058: that's interesting (S37K)

(6) 'Discursive' use:
S0278: he's a lovely fella ain't he?
S0013: course he is.
S0278: well thank you very much.
S0013: **that's okay** (S7RA)

In all cases, the pattern THAT'S ADJ is used in utterance-initial position or preceded by an interjection in this position. The demonstrative pronoun *that* refers to the content of one or more preceding utterances in what Halliday and Hasan (1976) call "extended anaphoric reference". The predicate consists of the contracted form of the copula and an adjective. In all cases, the communicative goal motivating speakers to use this pattern is the wish to relate back to something mentioned in the previous cotext and express some sort of comment.

The examples given illustrate the six most common specific functions of the pattern. In (1) the speaker offers a positive evaluation, in (2) a comment on the truth or correctness of what was said, and in (3) an assessment from an ethical perspective. Example (4) has a predominantly emotive function, and (5) a descriptive one. In example (6) the pattern is mainly used to signal uptake of what was said by the previous speaker, i.e. it has a predominantly discursive function. It should be emphasized that these six functions are idealized peaks in what is in fact a rather scattered pragmatic and semantic landscape. One utterance can be motivated by several goals and express a combination of, say, evaluation, epistemic confirmation and discursive uptake. Since many adjectives, e.g. *right* or *fine* or *lovely*, can be chosen to realize different functions in one or different utterances, there is no one-to-one correspondence between adjective types and functions and meanings. Nevertheless, given the programmatic nature of the study, we will pretend that such a one-to-one correspondence actually exists and shoehorn each adjective into the functional category that it instantiates most typically and frequently as indicated by the data.

The pattern THAT'S ADJ competes with a number of closely related patterns also offering the potential to combine extended anaphoric reference with various kinds of predications:

(7) THAT IS ADJ: e.g. *that is brilliant, that is interesting*
(8) THAT IS ADJ: e.g. *this is true, this is nice*
(9) IT'S ADJ: e.g. *it's weird, it's lovely.*
(10) THAT'S (A) N: e.g. *that's nonsense, that's a shame, that's a lie*
(11) WHAT A N: e.g. *what a shame, what a nightmare*

Even though these patterns clearly lie within the envelope of variation from an onomasiological perspective, they are not included in the present study. This restriction is necessary at this point to keep the methodological challenge within manageable bounds. Despite the fact that we are taking the form THAT'S ADJ as our point of departure, we conceive of our investigation as a study in onomasiological variation, because we focus on how different communicative goals are encoded by the choice of different adjectives. Semasiological variation, i.e., variation in the meanings of specific forms such as *that's right* or *that's fine* lies outside the scope of this study but should be included in future work.

DATA

Data Source

As the pattern under investigation is typically used in spontaneous spoken interaction, we decided to harvest data from the British National Corpus 2014 (BNC2014), which contains about 11 million words of transcribed casual conversations and has the additional advantage that metadata about speakers are available. BNC2014 is a successor to the British National Corpus (BNC).[1] So far only the spoken component, collected between 2012 and 2016, has been published. All words have been tagged with regard to both part-of-speech and, remarkably, semantic information, using the Lancaster UCREL English semantic tagger (USAS).[2] The corpus is also richly annotated with regard to various types of social and situational metadata.

Data Retrieval

Data from the BNC2014 can be accessed and queried online using the CQPweb interface provided by Lancaster University[3] or downloaded for use with individual processing methods. While the online platform offers a sophisticated interface to perform complex linguistic queries using the CQP query language, it does not enable users to include all metadata and to export results in a format that allows for more fine-grained analyses and filtering.

The freely available offline version of the BNC2014 provides all corpus texts with annotations in XML format as well as spreadsheets containing the full corpus metadata. Parsing this archive allows users to perform complex queries on the full textual data as well as to analyze hits according to conversation- and speaker-based information, which is essential for investigating variation in these two dimensions.

We therefore created a Python script that processes the BNC2014 data using XML parsing to enable queries based on all tags available in the textual data.[4] In the case of the pattern THAT'S ADJ we retrieved all utterances that either start with the pattern or where it is only preceded by interjections (POS: UH). Based on this restriction we collected all attestations that feature the singular determiner *that* (POS: DD1) followed by the item *'s* and an adjective (POS: JJ). In addition to collecting all instances of the pattern, the script outputs the total number of attestations in the corpus (n = 4,883).

An inspection of the results revealed a number of false positives, mainly stemming from tagging errors, which could be reduced by additional filtering using a blacklist of six tokens: *to, timing, news, bullshit, awesome, enough.*

For each hit we store the full attestation (e.g., *that's good*), the slot-filling adjective (e.g., *good*), its semantic tag (e.g., A5:1) and its category description (e.g., EVALUATION:- GOOD/BAD), which we add from the USAS tagset. Besides, we record utterance, conversation and speaker IDs which allow us to automatically retrieve all metadata from the BNC2014 spreadsheets and to include it in the output: e.g., AGE, GENDER, BIRTHPLACE of speakers or DATE, TOPIC and TYPE of conversations. Based on the list of speakers who have used the pattern we then query the full corpus to calculate the total number of words contributed by each individual, which is needed to determine normalized frequencies per speaker and to perform statistical tests targeting individual variation.

The Python script can be found in the supplementary material attached to this article. While this script has been tailored to detect instances of the THAT'S ADJ pattern, it can be readily adapted to perform any XML-based query in the BNC2014 by modifying only the query part of the script.

Manual Post-processing

Although the precision of the automatic processing was high, 268 false hits (amounting to 5.49%) had to be removed manually from the dataset. The major types of unwanted hits were: 1) uses of *that* which clearly functioned as relative rather than demonstrative pronouns (see 12); 2) uses of the pattern *THAT'S* ADJ with deictic reference to objects accessible in the situational context (see 13); or 3) with anaphoric reference to antecedents referring to concrete objects (see 14).

(12)
S0084: it's better to find something that you can do.
S0083: mm
S0084: **that's stable** in the short-term (.) and get a qualification that means you it will be stable rather than just here and there.

(13)
S0245: what color do you want?
S0246: grey

[1] http://corpora.lancs.ac.uk/bnc2014/
[2] http://ucrel.lancs.ac.uk/usas/
[3] https://cqpweb.lancs.ac.uk/

[4] Desagulier (2014) provides an R-script for extracting linguistic data and social metadata from the offline version of the BNC2014. The script allows users to perform basic queries using word and part-of-speech information and includes some metadata about speakers in its output. It also offers user-friendly options to create some exploratory plots and to export results to text files. However, the script does not provide options to formulate more complex queries, e.g. filtering based on position and tag sequences, or to retrieve and export semantic information and metadata about conversations and speakers required for this study.

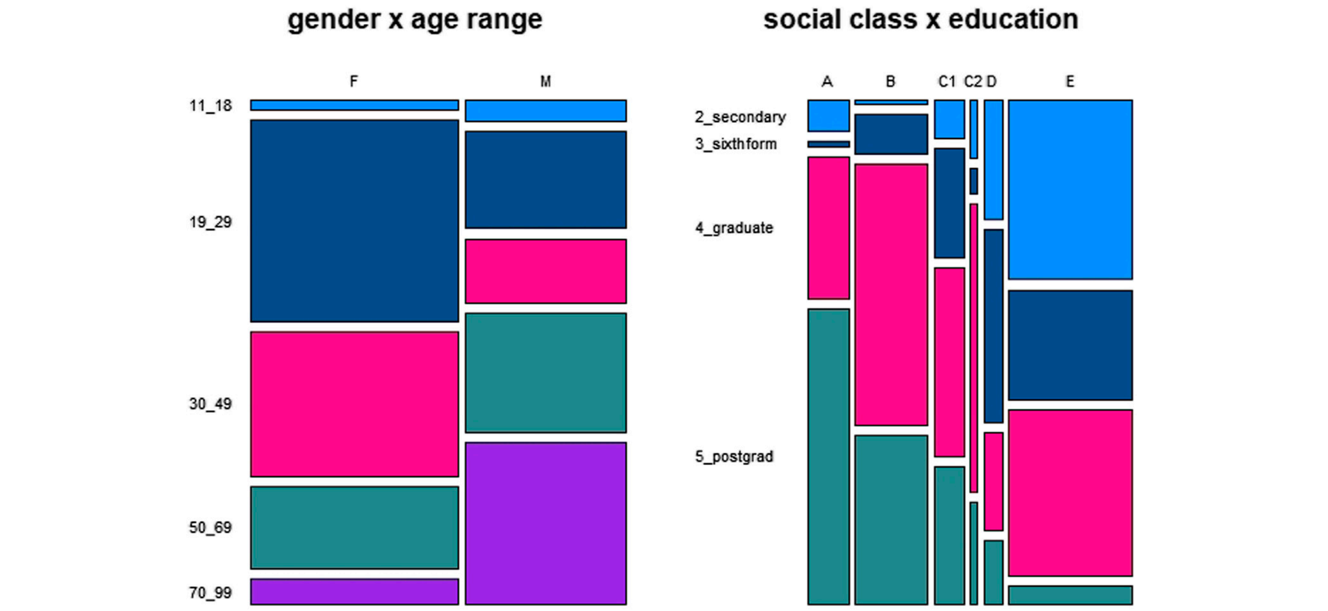


FIGURE 1 | Data distribution across the variables gender and age and social class and education.

S0245: I want grey shall we get two? it's only two fifty (.) comes in bla-
S0246: oh wait **that's black** (BNC2014, S4QK 92).

(14)
S0515: this is called the Lipstick Tower
S0512: oh uhu
S0515: **that's modern** (BNC2014, SGAW 465).

The dataset had to be adjusted in three more ways. First, all 23 attestations contributed by three speakers in the age range 0–10 were removed. Also removed were 208 attestations that featured the value "unknown" for one or more social variable. And third, due to data scarcity in some of the age ranges, we re-categorized the variable age into five instead of the original 10 age ranges, comprising ages 11 to 18, 19 to 29, 30 to 49, 50 to 69 and 70 to 99, respectively.

Final Dataset

The final dataset includes 4,394 attestations by 445 speakers in 931 conversations. These 4,394 tokens represent 159 adjective types. Boasting as many as 1,418 tokens, the most frequent adjective is *right*, followed by *good* (484 tokens) and *true* (340 tokens). 62 adjectives occur only once, 15 adjectives twice. The mean of tokens per adjective is 27.64, the median 3. The maximum number of tokens per speaker is 525, the minimum 1. The mean of tokens per speaker is 9.88, the median 3.

Data Distribution for Major Social Variables

Generally, the data are not distributed evenly across the categories of the social variables included in the metadata of BNC2014. We focus on the distribution of the main variables gender, age, education and social class. As is indicated by the mosaic plot given in **Figure 1**, there are more data by women than by men, more data by young women than by older women and more data by older men than by young ones. As far as education and social class are concerned, **Figure 1** indicates a substantial overrepresentation of social classes E and B and an expectable trend for a positive correlation between higher levels of education and social class.

RESULTS: DESCRIPTIVE DATA SUMMARY

Distribution of Tokens and Types Across Semantic Classes

As is shown in **Table 1**, there is no positive correlation between numbers of tokens and types. The class boasting the largest number of tokens, i.e., "epistemic", is on the second-lowest rank regarding types, while "descriptive" is manifested by the largest number of types and the second-lowest number of tokens. The class "uptake" stands out because it is only represented by the three types *alright, fine* and *good* (see also **Table 2**).

Most Frequent Adjectives Per Semantic Class

Table 2 lists the most frequent adjective types per semantic class. The frequency thresholds selected are provided in the header of

TABLE 1 | Distribution of tokens and types across semantic classes.

Class	Tokens	Types of adjectives
Epistemic	1853	9
Evaluative	1,177	33
Uptake	597	3
Emotive	472	37
Descriptive	468	63
Ethical	48	20
Total	4,615	165

TABLE 2 | Most frequent adjectives per semantic class.

Epistemic (all)	n	Evaluative (n > 9)	n	Uptake (all)	n	Emotive (n > 9)	n	Descriptive (n > 9)	n	Ethical (n > 1)	n
right	1,477	*good*	512	*alright*	277	*amazing*	103	*weird*	89	*fair*	11
true	350	*nice*	199	*fine*	224	*funny*	79	*interesting*	66	*harsh*	5
wrong	11	*cool*	130	*okay*	96	*ridiculous*	51	*crazy*	57	*poor*	4
correct	7	*brilliant*	63	—	—	*awful*	34	*different*	16	*mean*	4
impossible	3	*great*	58	—	—	*horrible*	28	*strange*	16	*nasty*	3
incorrect	2	*lovely*	44	—	—	*disgusting*	26	*clever*	15	*naughty*	3
exact	1	*terrible*	39	—	—	*awesome*	24	*cute*	13	*unfair*	2
definite	1	*bad*	37	—	—	*hilarious*	23	*mental*	12	*scandalous*	2
unlikely	1	*incredible*	15	—	—	*annoying*	15	*pretty*	12	*generous*	2
—	—	*fantastic*	14	—	—	*sad*	12	*stupid*	12	*vile*	2
—	—	*perfect*	12	—	—	*exciting*	11	*beautiful*	11	—	—
—	—	—	—	—	—	—	—	*mad*	10	—	—
—	—	—	—	—	—	—	—	*easy*	10	—	—

the table. It should be noted that the class of "epistemic" adjectives is strongly dominated by *right* and, to a much lesser degree, *true*, while the other classes show a much less steeply declining frequency distribution.

Distribution of Semantic Classes Across Social Variables

Figure 2 provides a survey of the frequency distribution of semantic classes across the four major social variables.

With regard to the variable GENDER, the proportions of the classes "ethical" and "uptake" are very similar for "female" and "male". "Epistemic" adjectives account for a higher proportion of the tokens of men than of those of women, which is made up by higher proportions for "descriptive", "emotive" and "evaluative" used by women.

Regarding AGE, we see very high proportions of "epistemic" adjectives for the age ranges 60 to 69 and 70 to 79. This corresponds to low proportions for the other classes in comparison to the younger age groups, who use adjectives of the other classes relatively more frequently.

The plot for the variable SOCIAL CLASS does not show a clear trend from class "A" to "E". Instead, there is a U-shaped pattern with "C1" and "C2" in the center and similar trends in both directions: an increase of "epistemic" and a decrease of "evaluative" toward "A" and "B" as well as "D" and "E".

The data for EDUCATION also do not reflect a consistent trend, but instead seem to indicate more or less random variation.

Distribution of the Twelve Most Frequent Adjectives across Social Variables

Figures 3-**6** zoom in on the 12 most frequently used adjectives and represent their distribution across the four major social variables. **Figure 3** representing the variable GENDER shows a more or less even distribution for the adjective *alright*, a strong male preponderance for *right*, and a female preponderance for the rest, which is particularly strong for the evaluative adjectives *amazing, funny, nice*.

Figure 4, rendering the data for the variable AGE, shows a general tendency for lower frequencies with higher age, in particular for *cool* and *weird*, and a reverse tendency for *right*.

As is indicated by **Figure 5**, the variable SOCIAL CLASS shows a clear trend for *right* to be used more frequently by members of higher social classes. Otherwise, there are no obvious tendencies. The same is true for the results regarding the variable EDUCATION, as shown by **Figure 6**.

Variation Across Conversations–Semantic Classes

Social variation is compromised and superseded by situational variation (Labov 1966). Given the structure of our dataset, a good way of describing the effect of situational variation is to look at variation across conversations. **Figure 7** represents the distribution regarding semantic classes in all 46 conversations which contain more than 15 instances of the target pattern. Overall, we notice a strong preponderance of the class "epistemic", which is mainly caused by the very high frequency of *right*. However, some conversations show a more distributed pattern, e.g. conversations S28F, S9P6, SM88, STWC, SU82 or SWWZ. The conversations S28F, S64H, S8PW and STWC are dominated by the use of "evaluative" adjectives, the conversations SFNQ and SKHW by "emotive" adjectives. Assessing the interaction between situational variation and social variation will be left to the inferential statistics reported in *Inferential statistics and results*, because it is too complex for descriptive techniques.

Individual Speaker Variation

Finally, we zoom in on differences between individual speakers, which are one of the main concerns of this paper. We will select different portions of the dataset, depending on how well they lend themselves to various ways of describing findings.

Speakers' Choice of Semantic Classes

For the description of speakers' choices of semantic classes, we have selected the data from the 30 speakers who boast a frequency of higher than 30 tokens. **Figure 8** shows their distributions, ordered in terms of the frequency of uses of the pattern. Overall, the figure indicates a very large degree

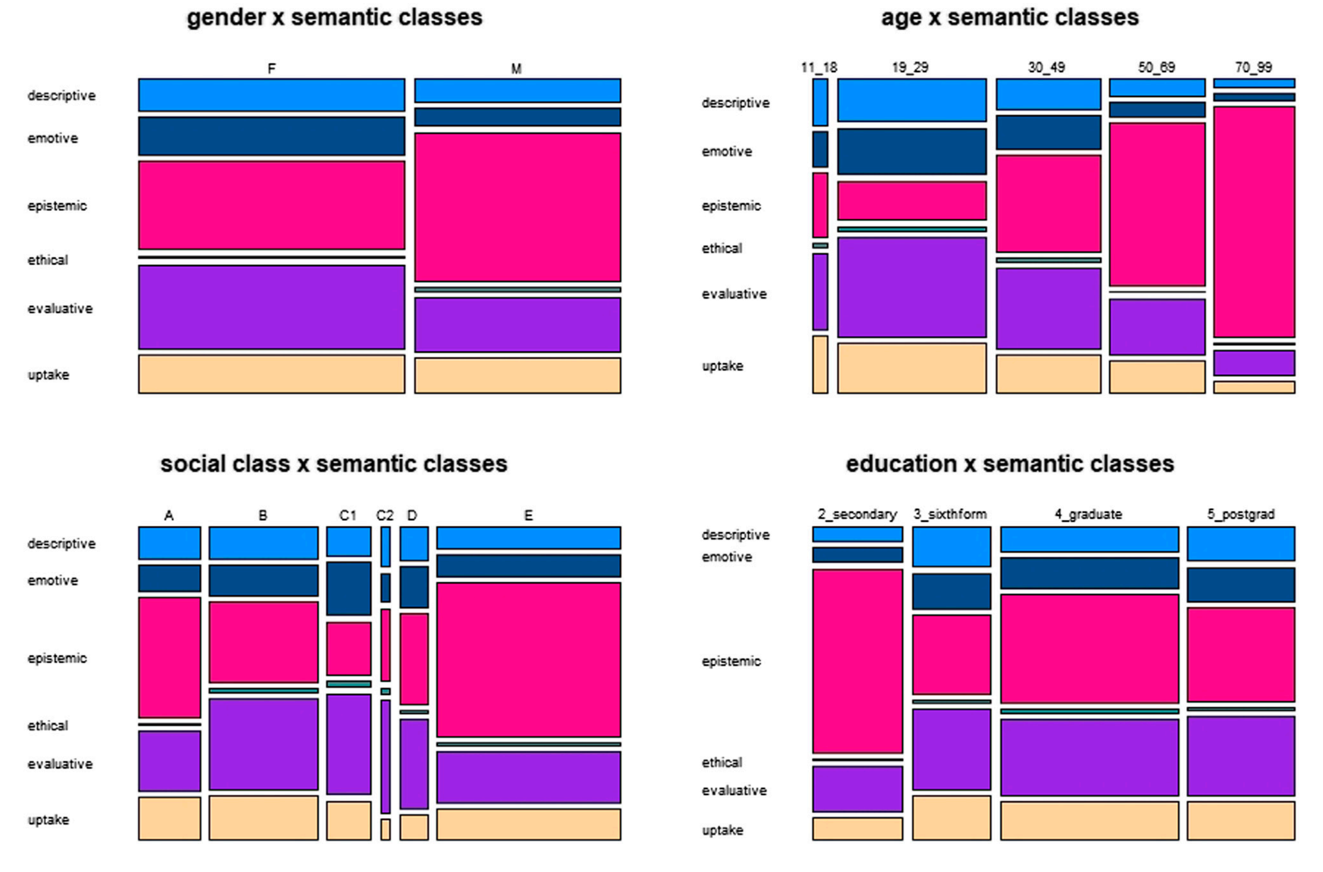


FIGURE 2 | Distribution of semantic classes across social variables.

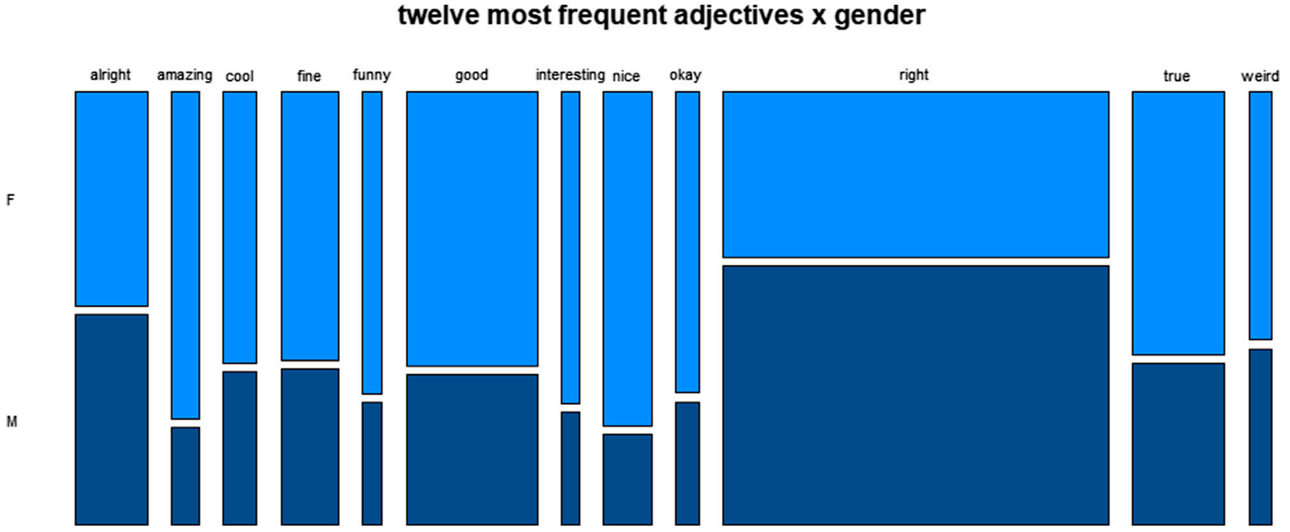


FIGURE 3 | Distribution of twelve most frequent adjectives across gender.

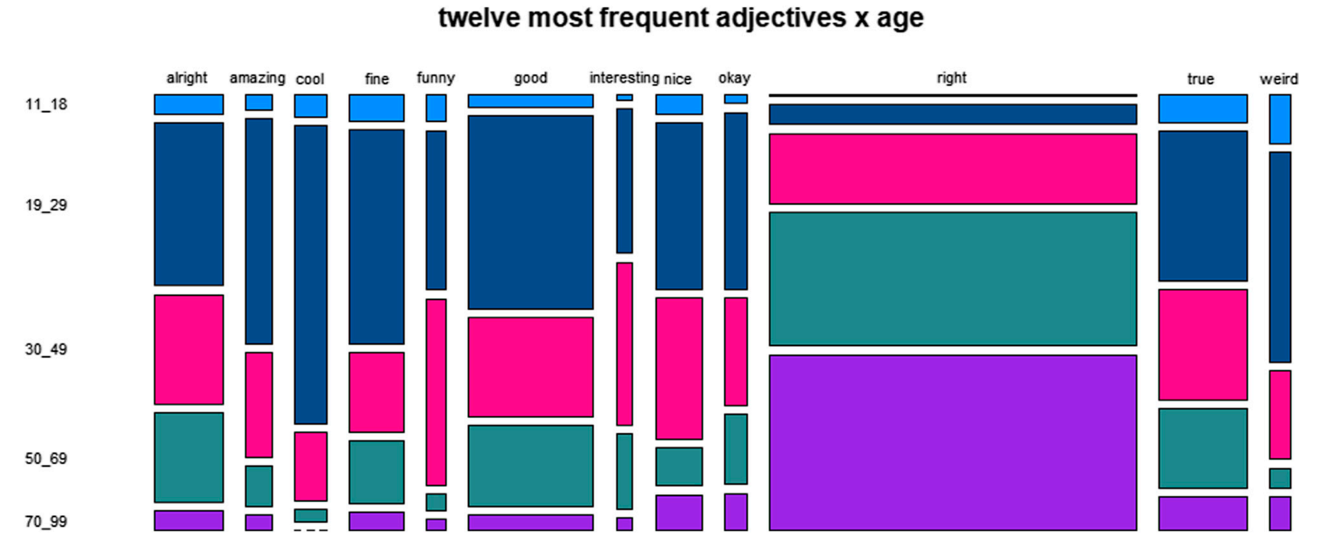


FIGURE 4 | Distribution of twelve most frequent adjectives across age.

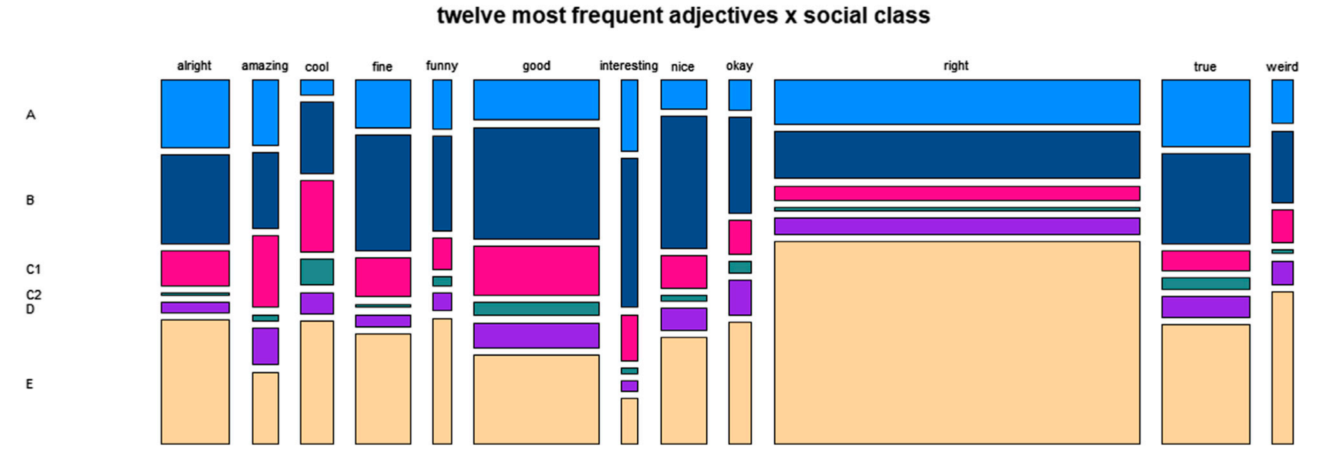


FIGURE 5 | Distribution of twelve most frequent adjectives across social class.

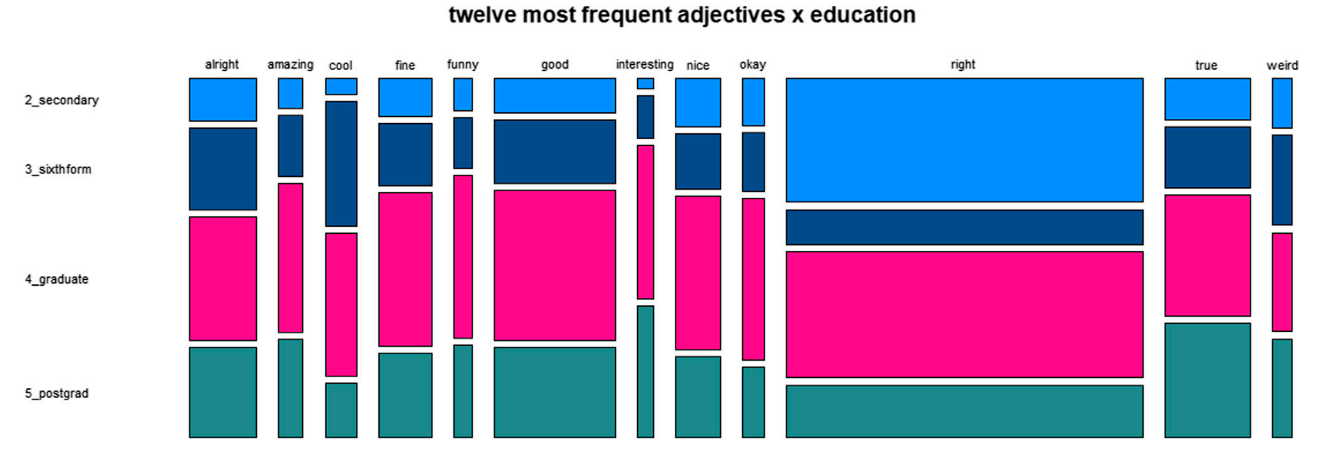


FIGURE 6 | Distribution of twelve most frequent adjectives across education.

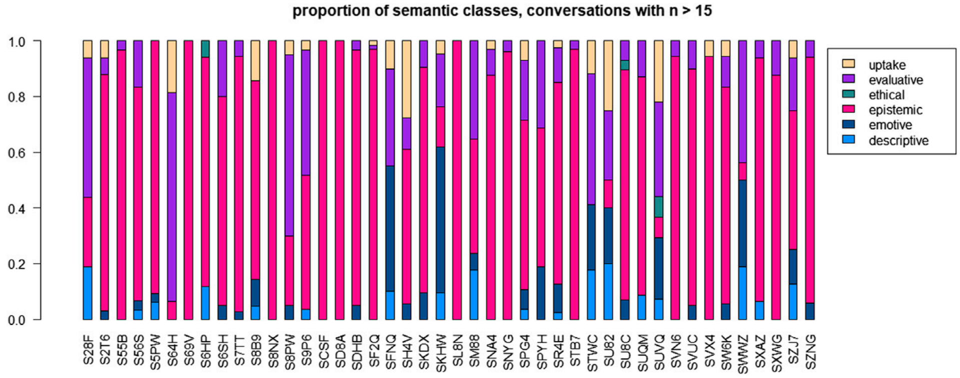


FIGURE 7 | Distribution across semantic classes for 46 conversations with n > 15.

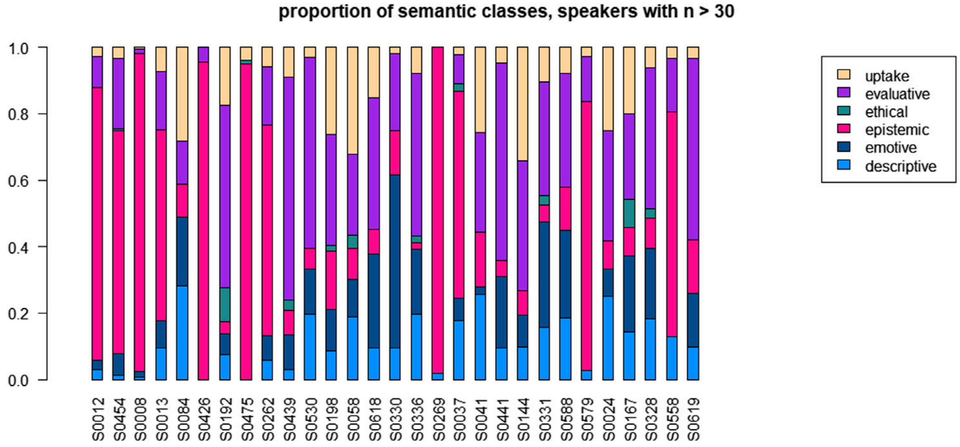


FIGURE 8 | Distribution across semantic classes for 16 speakers with n > 50.

of inter-individual variation. The figure allows the following observations:

- a dominance of "epistemic" for 10 speakers: S0012, S0454, S0008, S0013, S0426, S0475, S0262, S0269, S0037, S0579;
- dominance for "evaluative" for eight speakers: S0192, S0439, S0530, S0618, S0336, S0441, S0328, S0619;
- dominance of "uptake" for two speakers: S0058, S0144;
- dominance of "emotive" for one speaker: S0330;
- and a quite balanced distribution for nine speakers: S0084, S0198, S0618, S0525, S0041, S0331, S0588, S0024, S0167

Speakers' Choice of Specific Adjectives

As representing the data for speakers' choices of adjectives requires more space, we select the ten speakers with the largest number of tokens, from n = 525 to n = 68 (see **Figure 9**). Social characteristics are provided in the legend of all 10 panels of **Figure 9**. Not surprisingly, *right* turns out to be the dominant choice by far for as many as seven speakers (S0012, S0454, S0008, S0013, S0426, S0475, S0262). However, the degree of this dominance varies considerably from very extreme cases such as S0008 to more moderate ones such as S0454. What is also remarkable is that the slope of the curves outlined by the bars show very different shapes, reflecting the extent to which individual speakers favor only one or a small number of adjectives. In addition, it seems more or less impossible to correlate the differences between speakers with their social characteristics.

Figure 9 also provides the data for three speakers who do not have the routine of choosing *right* more frequently: S0084 favors the adjective *fine*, S0192 and S0439 the adjective *good*, followed by *cool* in both cases. Both of these speakers are young, as would be expected by the choice of *cool*, one is male and the other female.

Overall, the panels in **Figure 9** show a mixture of speakers with extreme habits PLUS a range of other adjectives (S0454, S0013) and speakers with extreme habits WITHOUT noteworthy frequencies of other adjectives (S0008, S0475). This is an important observation that we will come back to in *Social and cognitive implications* below.

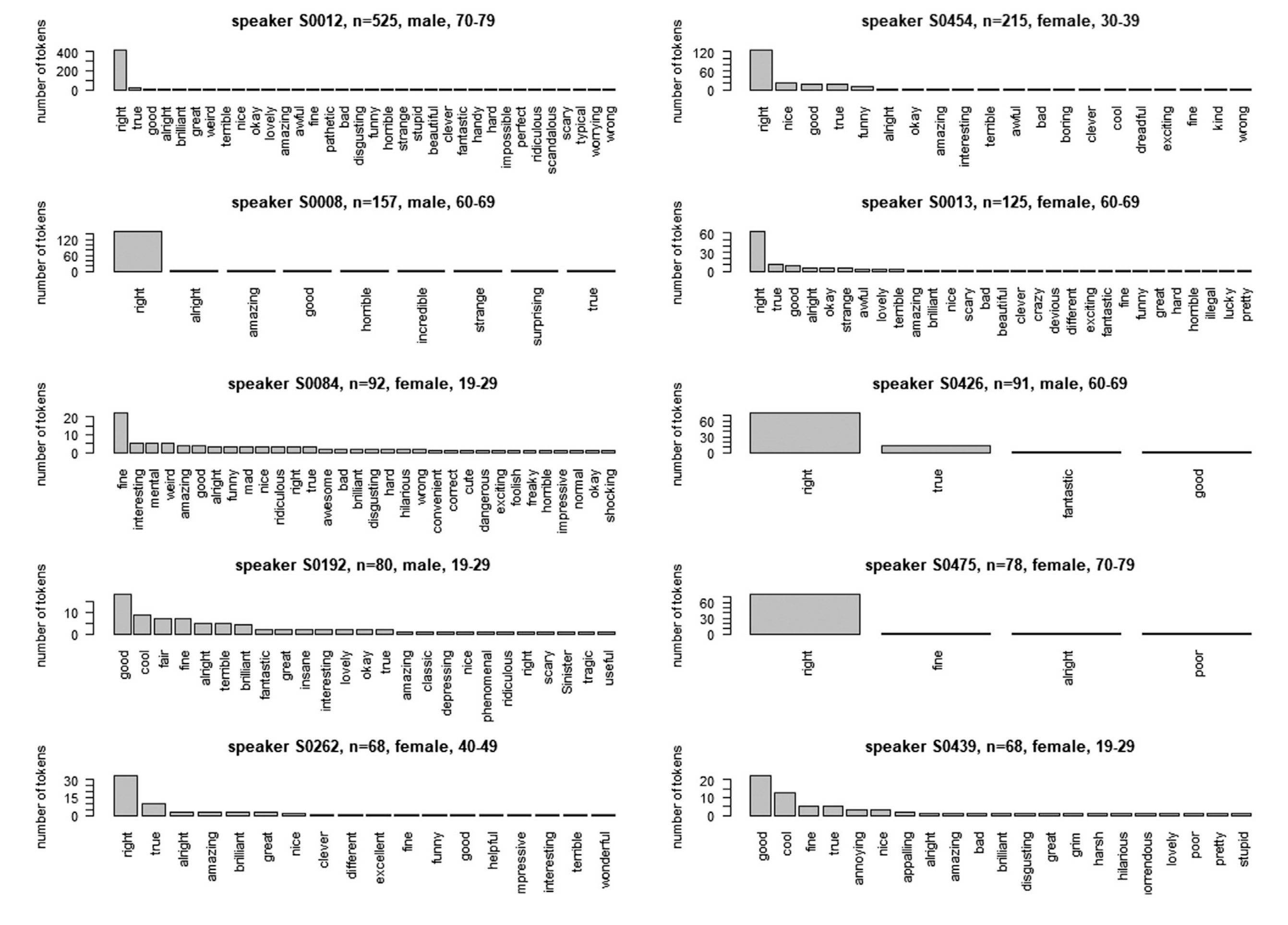


FIGURE 9 | Distribution of adjectives for 10 speakers with highest frequencies of the pattern.

INFERENTIAL STATISTICS AND RESULTS

Aims

The aim of the inferential statistics reported in this section is to model the effects of social, situational and speaker variation. Specifically, we want to gauge.

a. the effects of the four social variables GENDER, AGE, EDUCATION and SOCIAL CLASS on the choice of semantic classes and the most frequent adjectives;
b. the influence of situation-dependent variation by looking at the effects of the variable CONVERSATION;
c. the influence of individual variation by looking at the variable SPEAKER.

This will enable us to answer the question to what extent the variation found can be explained by social variables and to what extent it is superseded by situational and individual variation.

Statistical Models

To reach these goals, we fitted mixed-effects binomial logistic regression models using GENDER, AGE, EDUCATION and SOCIAL CLASS as fixed effects and SPEAKER and CONVERSATION as random ones. This was done using the glmer function of the lme4 package (Version 1.1–23) in R (Version 4.0.2). The inclusion of the two random effects allows us to gauge the extent to which variation can be attributed to differences between individual speakers or conversations. The random effect SPEAKER increases to the extent that individual speakers show a tendency to repeat their choices of semantic classes and adjectives, and the random effect CONVERSATION increases to the extent that choices are repeated by the participants within one conversation. It is in this way that the random effect SPEAKER can be interpreted as an indicator of individual habits and inter-individual variation, and the random effect CONVERSATION as an indicator of same-speaker and other-speaker repetition in conversations.

The models targeted dependent variables on the two levels of analysis also used for the descriptive statistical analysis: the choice of semantic classes and the choice of specific adjectives. With regard to semantic classes, the binomial models compare the choice of one semantic class (e.g., “epistemic”) to all instances of all other classes. With regard to adjectives, we compare selected adjectives to all semantically similar adjectives from the same semantic class (see *Choice of*

TABLE 3 | Results from the mixed-effects binomial logistic regression model. The outcome variable was the use of an "epistemic" adjective. Random effects for CONVERSATION and SPEAKER were included.

	Epistemic		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	−1.87		
Gender [M]	−0.05	−0.48–0.37	0.801
Age [19_29]	−0.67	−1.62–0.28	0.166
Age [30_49]	−0.10	−1.11–0.91	0.848
Age [50_69]	1.22	0.27–2.16	0.012
Age [70_99]	2.92	1.84–4.01	<0.001
Social Class [B]	−0.58	−1.24–0.08	0.086
Social Class [C1]	−0.29	−1.08–0.50	0.468
Social Class [C2]	0.04	−1.36–1.44	0.952
Social Class [D]	0.42	−0.49–1.33	0.368
Social Class [E]	0.14	−0.53–0.81	0.685
Education [3_sixthform]	0.09	−0.64–0.82	0.807
Education [4_graduate]	0.38	−0.32–1.08	0.288
Education [5_postgrad]	0.58	−0.18–1.35	0.136
Random Effects			
σ^2 conversation	0.74	—	—
σ^2 speaker	1.37	—	—
ICC conversation	0.09	—	—
ICC speaker	0.33	—	—
Number of observations			
Observations	4,394	—	—
N conversation	931	—	—
N speaker	445	—	—

Specific Adjectives for more details). This corresponds to the conceptually plausible assumption that a speaker planning to use the pattern THAT'S ADJ has a twofold paradigmatic choice between the general types of meaning they want to encode, on the one hand, and the specific adjective they want to use in order to do so, on the other.

Results

Results will be reported in two steps: *Choice of semantic classes* deals with regression models targeting the choice of semantic classes, and *Choice of specific adjectives* with those targeting the choice of specific adjectives.

Choice of Semantic Classes

Mixed-effects regression models were fitted for the three semantic classes boasting the highest frequencies of tokens, i.e., "epistemic", "evaluative" and "uptake". In all cases, we fitted two models: one based on the full dataset and one in which all speakers who contributed only one token were excluded. Since the results of the two models were very similar, we will only report those based on the full dataset.

In **Table 3**, we present the summary of the regression model for the choice of the semantic class "epistemic". The base categories are GENDER "female", AGE "11 to 18", SOCIAL CLASS "A" and EDUCATION 'secondary'. The summary indicates that the only social variable that was found to be significant was AGE, with speakers in the age ranges 50 to 69 and 70 to 99 showing a significantly incidence of using this semantic class. This is in line with expectations derived from the descriptive statistics reported in *Distribution of semantic classes across social variables*.

The two random effects CONVERSATION and SPEAKER can be gleaned from the standard deviations reported in the summary, which are 0.74 and 1.37, respectively. Especially for the variable SPEAKER, the score indicates a very strong effect of the repeated choices of individual speakers. A possible way of gauging the proportion of stochastic variation contributed by the random effects is to use the intra-class correlation coefficient (ICC; Nakagawa and Schielzeth 2010). This coefficient measures the correlation between the variance of a given random effect and the total variance. It is calculated by dividing the variance of a given random effect by the total random variation, i.e. the sum of the variance of all random effects and the variance of the logistic distribution. Since the latent-scale distribution-specific variance for the logit models we are using here is a constant given as $\pi^2/3$ (Nakagawa and Schielzeth 2010), the ICC for the random variable SPEAKER, for example, can be calculated as $\sigma^2_{\mathrm{SPEAKER}}/(\sigma^2_{\mathrm{SPEAKER}} + \sigma^2_{\mathrm{CONVERSATION}} + \pi^2/3)$. The ICCs for the random effects SPEAKER and CONVERSATION are 0.33 and 0.10, respectively. This can be interpreted as indicating that a proportion of 33% of the stochastic variation on the latent scale is contributed by the variable SPEAKER, and 10% by CONVERSATION.

Table 4 reports the summary of the regression model for the class "evaluative". The model indicates a weak but significant positive effect for SOCIAL CLASS "B" and a strong and highly significant negative effect for the age range 70–99. The variance rendered for each of the random effects are lower than for the class "epistemic"–0.40 for CONVERSATION and 0.64 for SPEAKER. However, with an ICC of 15%, the contribution of the variable SPEAKER to the stochastic variation remains considerable ($\mathrm{ICC}_{\mathrm{CONVERSATION}}$ = 10%).

TABLE 4 | Results from the mixed-effects binomial logistic regression model. The outcome variable was the use of an "evaluative" adjective. Random effects for CONVERSATION and SPEAKER were included.

	Evaluative		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	−0.87		
Gender [M]	−0.20	−0.50–0.10	0.196
Age [19_29]	0.03	−0.65–0.70	0.942
Age [30_49]	−0.04	−0.76–0.67	0.903
Age [50_69]	−0.36	−1.05–0.32	0.297
Age [70_99]	−1.55	−2.41–−0.69	<0.001
Social Class [B]	0.49	0.03–0.96	0.037
Social Class [C1]	0.41	−0.14–0.96	0.140
Social Class [C2]	0.77	−0.15–1.69	0.102
Social Class [D]	0.22	−0.42–0.86	0.502
Social Class [E]	−0.09	−0.57–0.40	0.730
Education [3_sixthform]	−0.28	−0.82–0.25	0.296
Education [4_graduate]	−0.25	−0.77–0.27	0.343
Education [5_postgrad]	−0.28	−1.70–−0.03	0.329
Random Effects			
σ^2 conversation	0.63	—	—
σ^2 speaker	0.80	—	—
ICC conversation	0.10	—	—
ICC speaker	0.15	—	—
Number of observations			
Observations	4,394	—	—
N conversation	931	—	—
N speaker	445	—	—

TABLE 5 | Results from the mixed-effects binomial logistic regression model. The outcome variable was the use of an "uptake" adjective. Random effects for CONVERSATION and SPEAKER were included.

	Uptake		
Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	−1.93		
Gender [M]	0.33	−0.01–0.68	0.056
Age [19_29]	0.38	−0.39–1.15	0.330
Age [30_49]	0.23	−0.60–1.06	0.586
Age [50_69]	−0.11	−0.89–0.67	0.778
Age [70_99]	−1.22	−2.18–0.26	0.013
Social Class [B]	0.21	−0.33–0.74	0.445
Social Class [C1]	−0.14	−0.79–0.50	0.669
Social Class [C2]	−1.10	−2.34–0.13	0.080
Social Class [D]	−0.66	−1.45–0.13	0.100
Social Class [E]	0.10	−0.46–0.66	0.725
Education [3_sixthform]	−0.15	−0.76–0.46	0.632
Education [4_graduate]	−0.36	−0.96–0.23	0.228
Education [5_postgrad]	−0.54	−1.19–0.12	0.108
Random Effects			
σ^2 $_{conversation}$	0.82	—	—
σ^2 $_{speaker}$	0.85	—	—
ICC $_{conversation}$	0.14	—	—
ICC $_{speaker}$	0.15	—	—
Number of observations			
Observations	4,394	—	—
N $_{conversation}$	931	—	—
N $_{speaker}$	445	—	—

TABLE 6 | Results summary for mixed-effects logistic regression models for *right, true, good, nice, alright* and *fine.* Random effects for CONVERSATION and SPEAKER were included.

Adjective	Significant fixed effects (estimate, significance level)	Random effects (standard deviation)	ICCs
right	Compared to all other epistemic adjectives meaning "true, correct"		
	AGE [30_49]: 2.85*	CONVERSATION: 1.22	15%
	AGE [50_69]: 4.51***	SPEAKER: 2.24	51%
	AGE [70_99]: 6.50***		
true	Compared to all other epistemic adjectives meaning "true, correct"		
	AGE [30_49]: -3.44**	CONVERSATION: 1.20	14%
	AGE [50_69]: -4.88***	SPEAKER: 2.39	55%
	AGE [70_99]: -6.74***		—
good	Compared to all other positive evaluative adjectives		
	—	CONVERSATION: 0.41	4%
	—	SPEAKER: 0.66	11%
nice	Compared to all other positive evaluative adjectives		
	GENDER [M]: -0.72*	CONVERSATION: 0.85	16%
	—	SPEAKER: 0.74	12%
alright	Compared to all adjectives		
	GENDER [M]: 0.69**	CONVERSATION: 1.15	24%
	AGE [70_99]: -1.42*	SPEAKER: 0.98	17%
	SOCIAL CLASS [D]: -1.36*		
fine	Compared to all adjectives		
	AGE [70_99]: -1.38*	CONVERSATION: 0.84	13%
		SPEAKER: 1.15	25%

The regression model for the semantic class "uptake" reported in **Table 5** indicates almost equally strong effects of the two random variables, with ICCs amounting to 14% for conversation and 15% for speaker. The only relevant social predictor is again AGE, with a weak but significant decrease associated with the age range 70–99.

In sum, the regression models suggest that effects of the fixed social variables on the choice of semantic classes are limited for all three semantic classes, while those of the random effects CONVERSATION and especially SPEAKER are considerable throughout and very strong for the semantic class "epistemic".

Choice of Specific Adjectives

The two top-ranking adjectives from the classes "epistemic", "evaluative" and "uptake" were selected for regression models targeting the choice of specific adjectives: *right* and *true*, *good* and *nice*, and *alright* and *fine*, respectively. We fitted models that compared these adjectives to all quasi-synonymous adjectives of the same class. For example, *right* was compared to all other epistemic adjectives with the meaning "true, correct", i.e., *true, correct, definite* and *exact*. *Good* was compared to the 18 other positive evaluative adjectives, including *brilliant, cool, excellent, fantastic, great* and *lovely*. This corresponds to the assumption that speakers select the adjectives from the pool of all those that can be used in the pattern in a given context. Since the group of uptaking adjectives includes no more than three adjectives, i.e., *alright, fine* and *okay*, the two adjectives *alright* and *fine* were compared to all other adjectives.

Rather than rendering the complete summaries of the regression models for all six adjectives, we restrict ourselves to reporting fixed effects that are significant at 5% level (stated as estimates and indicators of significance levels), random effects (stated as standard deviations) and ICCs per adjectives. This is summarized in **Table 6**.

The two epistemic adjectives *right* and *true* show opposite trends regarding the variable AGE, with *right* being favored with increasing age and *true* being disfavored. These effects are huge. AGE is also a relevant variable for the choice of the two "uptake" adjectives *alright* and *fine*. "Male" GENDER has a reducing effect on the choice of *nice* and an increasing one on the choice of *all right*. SOCIAL CLASS "B" has a reducing effect on *right*, and SOCIAL CLASS "D" also a reducing one on *alright*. Overall, the amount of variation that can be explained with the help of fixed social variables is astonishingly low, except for AGE with respect to *right* and *true*. In contrast, as in the case of the choice of semantic class, the two random variables SPEAKER and CONVERSATION show strong effects on the choice of adjectives. In all cases except *nice* and *alright*, the effect for SPEAKER is much stronger than that for CONVERSATION. *Right* and *true* stand out with stunningly high ICC scores in addition to the large effects for AGE, which suggest that the dominant factors determining the choice of these two adjectives in the pattern are speakers' habits–observable within and across conversations–and self- and other-repetition in conversations.[5]

[5] Recalling the quite extreme preference of speakers S0012, S0454 and S0008 reported for *right* in *Speakers' choice of specific adjectives*, one might assume that these three speakers are mainly responsible for the effects of AGE and SPEAKER. Therefore we also fitted models in which these three speakers were excluded, but the effects of AGE and SPEAKER remained almost equally large.

DISCUSSION

In this section we will first summarize the findings. These will then be discussed with regard to their social and cognitive implications. Finally, we will examine the relevance of these implications for the study of language variation and change. Throughout, we will take the perspective of the so-called *Entrenchment-and-Conventionalization Model* (Schmid 2020). We consider this model to be particularly suited for explaining the findings, because it integrates linguistic usage patterns, their conventionalization in the community and their entrenchment in the minds of individuals and tries to explain how the interaction between these three elements controls language structure, variation and change.

Summary of Findings

Speakers can use a wide range of adjectives in the lexico-grammatical pattern THAT'S ADJ in order to express various meanings (which can also be combined in specific utterances). We investigated 4,394 tokens of this pattern retrieved from BNC2014, originally produced by 445 speakers using 159 adjective types in 931 conversations. The descriptive and inferential statistics presented in this paper converge in the following findings:

- Speakers vary very strongly with regard to the frequency with which they a) use the pattern, b) encode the different meanings, and c) use the different adjectives.
- Overall, the effects of the social variables on the observed frequencies were fairly limited: higher AGE was found to have an increasing effect on the class of "epistemic" adjectives and the choice of the most frequent adjective *right*, and a decreasing effect on "evaluative" and "uptake" adjectives as well as the epistemic adjective *true*. GENDER influenced the choice of *nice* and *alright*, SOCIAL CLASS the choice of the semantic class "evaluative" and the adjectives *right* and *alright*.
- The effect of situational variation–approached via the random variable CONVERSATION–was found to be high throughout.
- Confirming the results of the descriptive statistical analysis, individual variation–approached via the random variable SPEAKER–was also found to have very strong effects on the choices of semantic classes and the adjectives focused on, with the class "epistemic" and the adjectives *right* and *true* standing out with extremely high effects of speaker repetition.

Social and Cognitive Implications

From the social perspective of the speech community, the sequence THAT'S ADJ qualifies as a highly conventionalized lexico-grammatical pattern whose use is motivated by a range of communicative functions. This means that among the members of the speech community, there is a mutually expected onomasiological regularity linking the goal to encode the various meanings of the pattern to its form and specific variants. Looking at the aggregated frequency distribution reported in *Distribution of tokens and types across semantic classes* and *Most frequent adjectives per semantic class*, one gets the impression that the pattern as such, its semantic variants and its specific instances such as *that's right, that's true* or *that's good* are indeed widely agreed upon means of reaching recurrent communicative goals. This is basically what is meant when we call the pattern *conventional*.

In general, this impression is certainly correct, but the aggregated macro-perspective glosses over the considerable variation found regarding the frequencies of choices of semantic classes and specific adjectives in different conversations and by individual speakers. From this perspective, the behavior of the speakers in the corpus turns out to be all but uniform.

How can these findings be explained? With regard to situational variation, there is a range of well-established factors that are likely to cause the effects observed for CONVERSATION: classic situational factors such as participants, setting, activity type, topic and register readily come to mind here. These could easily be looked at in greater detail, because a lot of the information that is required is available in the BNC2014 metadata.

In addition, and from a more cognitive and psycholinguistic perspective, one can assume that the participants involved in a conversation show the well-known tendency to repeat identical or semantically similar tokens of the pattern. This tendency has been described in terms of concepts such as *accommodation* (Giles et al., 1991; Giles and Ogay 2007), *alignment* (Pickering and Garrod 2004), *co-adaptation* (Larsen-Freeman and Cameron 2008; Schmid 2020), *dialogic resonance* (Du Bois 2014), *priming* (Pickering and Ferreira 2008) or *persistence* (Bock 1986). These notions can be invoked to explain the strong effects of the random variable CONVERSATION, because the interpersonal and psychological tendencies they refer to are reflected in the repetition of semantic classes and specific adjectives in the course of individual conversations.

As far as individual variation is concerned, the results concerning speakers' choices of adjectives (reported in *Individual speaker variation*) and those concerning the random variable SPEAKER in the mixed-effects regression models (see *Choice of specific adjectives*) indicate two things: first, that many speakers have routinized habits of using specific patterns such as *that's right*, *that's true* or *that's fine*; and second, that speakers' habits differ considerably and in ways that are not, or only weakly, determined by their social characteristics. It is true that in *Speakers' choice of specific adjectives* we found that many speakers showed a strong preference for the pattern *that's right*. But it is equally true that others hardly ever used this pattern and instead showed a high proportion of uses of *that's true* or *that's fine* or *that's good* in their data.

These findings can be interpreted from a cognitive perspective, if one accepts the logic of frequency-driven entrenchment (Langacker 1987; Schmid 2007). The premise of this logic is that what has become more entrenched by frequent repetition is activated more effortlessly and more quickly than what is less entrenched due to less frequent processing (Schmid 2017a; Schmid 2017b). If this premise is correct and if we reverse the perspective, one can assume that if a linguistic element or pattern is produced by a given speaker more frequently than another one which competes to encode the same information (Geeraerts 2017), then this element or pattern is more strongly entrenched in the mind of this speaker than the competing elements relative to the communicative task at hand (Schmid 2020). For example, one would assume that the pattern *that's right* is strongly entrenched in the mind of speaker S0012, who uses this pattern as many as 407 times in the BNC2014 data, with the quasi-synonymous pattern *that's true* trailing behind in second rank with as few as 21 instances (see **Figure 9** in *Speakers' choice of specific adjectives*). In contrast, speaker S0084 seems to have a particularly strongly entrenched representation of the pattern *that's fine*, and speaker S0439 of the pattern *that's good*.

It is tempting to claim that these strongly entrenched specific patterns are represented as holistic chunks in the minds of the respective speakers (Sinclair 1991; Wray 2002; Nelson 2018). Rather than putting together *that's* and *right* or *that's* and *fine* compositionally by means of syntactic operations, speakers who routinely use these patterns probably have them available as ready-made chunks or prefabs in their mental lexicons (Gobet et al., 2001; Ellis 2017). However, it is unclear how many repetitions are required to create such a chunk in the mental lexicon, and also, from a methodological point of view, how many attestations would be required as evidence for the existence of such a chunk (Blumenthal-Dramé 2012; Blumenthal-Dramé 2017). Therefore, following the arguments put forward by Schmid (2020), we argue that the chunk-like processing and representation of sequences is best accounted for in terms of particularly strong syntagmatic associations giving rise to a very high sequential predictability. In this way, more or less frequent patterns do not have to be forced into a categorical distinction between "chunk" and "compositional sequence", but can instead be described on a scale of strength of syntagmatic associations, from extremely strong and therefore essentially chunk-like to somewhat looser, as in the case of collocations or complementation patterns. The strength of syntagmatic associations is not only determined by the frequency of earlier processing episodes, but also by symbolic, paradigmatic and pragmatic associations. Symbolic associations connect the forms of the pattern to the various meanings. Paradigmatic associations connect the competitors in a given variable slot (e.g. *right* and *true* in the adjective slot of the pattern). Pragmatic associations connect the forms to communicative motivations and goals such as "express approval" or "express uptake" (Schmid 2014). From this perspective, the use of the pattern and its specific variants is not modeled as an either holistic or compositional access-retrieve-combine operation, but instead as the incremental activation of a dynamic pattern of the four types of association. In line with theories of predictive coding (Friston 2010; Huang and Rao 2011; Kroczek and Gunter 2017), this model of processing links representations based on prior experience with processing based on current perception and context.

An additional advantage of this associative approach is that it provides an integrated perspective on interesting differences between the "usage profiles" (Schmid and Mantlik 2015) of different speakers (see again *Speakers' choice of specific adjectives*). Speaker S0008, for example, is a very extreme case: the 157 tokens of the variable pattern THAT'S ADJ that he contributes to the corpus are divided into as many as 149 tokens of *that's right* and only one token each of *alright, amazing, good, horrible, incredible, strange, surprising* and *true*. The highly routinized repetition of *that's right* can be modeled as being triggered by an associative complex connecting the communicative goal of expressing consent and approval by means of the sequence *that's right*. This sequence seems to be so strongly entrenched pragmatically, symbolically and syntagmatically in the mind of this speaker that it does not seem to have any serious paradigmatic competitors for reaching the given communicative goal. This is presumably different in the case of speaker S0454, who also has a large proportion of uses of *that's right* (n = 124), but contributes another 91 tokens, among them 22 tokens of *nice*, 19 of *true* and *good* and nine of *funny*. This distribution can be interpreted as reflecting the co-existence of a strong specific representation of the syntagmatic sequence *that's right* (which is strongly triggered by pragmatic associations) and an entrenched variable pattern which is also connected to the function "evaluative" in addition to "epistemic". Further illuminating examples are speakers S0084 and S0192, whose use of the pattern is apparently dominated by several pragmatic motivations, including "uptake" and "evaluative", as is indicated by the frequent use of the adjectives *fine* as well as *interesting, mental, weird, amazing* and *good* in the case of S0084 and *good, cool, fair, fine, alright, brilliant* and *terrible* in the case of S0192.

Translating these claims into the more established but also more rigid and less dynamic framework of Construction Grammar (Goldberg 1995; Goldberg 2006; Goldberg 2019; Hilpert 2014)–which has been the elephant in the room anyway –, we can say that speakers differ with regard to how strongly the highly schematic THAT'S ADJ construction or lower-level schemas like "epistemic" or "evaluative" or certain lexically-specific constructions such as *that's right* or *that's fine* are represented in their constructicons.

In sum, we have claimed that underneath the apparent uniform linguistic behavior on the aggregate macro-level of the community we find significant differences in the frequencies of usage patterns, and that these differences can be interpreted as indicating a considerable degree of covert "speaker-specific cognitive variation" (Schmid 2020: 308). In the next, penultimate section we will discuss the ways in which individual differences and covert cognitive variation can affect variation and change on the macro-level and why they should be of interest to sociolinguists and students of language change.

Implications for the Study of Variation and Change

Variation

In sociolinguistics and language change, language has traditionally been framed as an "object possessing orderly heterogeneity" (Weinreich et al., 1968: 100), with *orderly* essentially referring to a differentiation of the behavior of groups of speakers which is systematic in the sense that it can be correlated with social and situational factors. Associated with the variables GENDER, AGE, EDUCATION and SOCIAL CLASS, in the present case study this orderly type of variation turned out to be less dominant than individual variation, which is unorderly by definition. Individual differences turned out to contribute much more to the overall variation observed than linguists on the hunt for orderly heterogeneity are usually happy to see.

One way out of this dilemma would be to state that the frequency distributions found are no more than what we have described them as, i.e., expectable effects of cognitive processes like entrenchment and priming. As such, one could conclude, they do not have anything interesting to contribute to our understanding of social variation and language change. However, this might be too easy a way out. After all, it is generally assumed in quantitative sociolinguistics and historical linguistics that frequency distributions reflect and reinforce sociolinguistic patterns and that differences in usage frequencies can trigger and index language change. This would suggest that individual frequency differences should not be ignored in the study of social variation and language change, but instead by related to social and situational variation.

How can this be achieved? What are the links between individual and social variation? In our view, individual variation neither compromises nor supersedes social variation, but rather generally subserves it. The fundamental assumption underlying sociolinguistics is that the entrenched routines and habits of speakers on all levels of language–from phonology and morphosyntax to pragmatics–are influenced by social factors or at least correlate with them. These social factors include the usual suspects, for example frequency of social interaction, the structure and density of social networks and communities of practice, people's tendency to seek solidarity and signal distance and to identify and align with members of their social groups and networks. Both orderly social and seemingly random individual variation are ultimately based on the routines and habits of speakers. Variation is considered to be orderly to the extent that these routines and the differences between them are correlated with some aspects of social structure or situated social interaction. In the *Entrenchment-and-Conventionalization Model* (Schmid 2020), it is generally assumed that speakers' patterns of social interactions and their social identities ultimately do shape the associative networks in their minds, because they determine the linguistic experiences that speakers accumulate. However, there seems to be a considerable residue of individual habits and whims which mainly have a cognitive foundation in the repetition-driven routinization of past behavior. It would therefore not be surprising if a closer look at existing quantitative sociolinguistic studies revealed that in many cases the usage patterns of individual speakers were a central source of variation to be taken much more seriously.

Change

The claim that individual variation should attract more attention gains further weight when we consider that the behavior of individual speakers can trigger and support various types of language change. The most obvious way in which this can happen is the use and subsequent repetition of new fillers of variable slots of existing patterns (Schmid 2020: 137). For example, tracing back the use of THAT'S ADJ and THAT IS ADJ in the Early English Books Online corpus (Petré 2016), one finds that for a considerable time after the pattern seems to have been borrowed from French around 1,500, only the epistemic adjectives *true* and *false* were used, with *right* not appearing before the middle of the 17th century. Descriptive and evaluative adjectives such as *good, excellent* or *strange* entered the scene around Shakespeare's time. These innovations must have been introduced by individual speakers, and their propagation was presumably supported by repeated use by a small number of speakers to begin with. Concrete illustrations of how this works in the case of other patterns can be found in Schmid and Mantlik (2015) and Mantlik and Schmid (2018).

High frequencies of repetition of specific sequences such as *that's right* by individual speakers also have the potential to trigger and support macro-changes like pragmaticalization (Diewald 2011) and grammaticalization (see Schmid 2020: Ch. 19 for discussion). In fact, *that's right* can be considered a case in point if one argues that–especially for those speakers who repeat this sequence very frequently–it is no longer an expression of epistemic stance, signaling a truth-related token of agreement, but has turned into a generalized discourse marker essentially on a par with the "uptake" adjectives *alright, fine* and *okay*. *Fine*, too, can be claimed to have undergone a similar pragmaticalization process from expressing an evaluative and hence propositional meaning to mainly serving a discursive function. Recent studies on the contribution of individual differences in language change–e.g. by Schmid and Mantlik (2015); Baxter and Croft (2016); Petré and Van de Velde (2018); Anthonissen (2020a); Anthonissen (2020b); Fonteyn and Nini (2020); Petré and Anthonissen (2020)–are accumulating more and more evidence suggesting that especially the early phases of the propagation of innovations are marked by massive variation among speakers, with some using a new element or pattern highly frequently while many contemporary writers do not use it at all (Schmid 2020: 320).

CONCLUSION

The explicit mission of quantitative variationist sociolinguistics has been–and will continue to be–to unveil sociolinguistic patterns, i.e., to identify correlations of types of linguistic behavior with types of speakers and types of situations. Individual differences have been considered an unwelcome, uninteresting and largely uncontrollable source of variation in this endeavor. Therefore, with notable exceptions (see e.g., Tagliamonte and Baayen 2012), researchers in this field have tended not to pay much attention to the effect of individual variation, even if speakers or test participants were included as random effects in mixed-effects models or random forests. Against this backdrop, the main thrust of this paper is of a theoretical and methodological nature, rather than related to the content in terms of subject-matter. We have argued that the study of individual variation should complement the study of social (including regional and situational) variation, mainly because individual variation ultimately subserves social variation and because it plays an important role in language change. The suggestions we have made as to how the study of individual differences can be approached are just a starting-point. They are meant to encourage scholars working in quantitative and especially computational sociolinguistics to step up their efforts to take individual variation on board in future work and to develop more sophisticated tools and techniques for investigating it.

DATA AVAILABILITY STATEMENT

Publicly available datasets retrieved from http://corpora.lancs.ac.uk/bnc2014/ were analyzed in this study. The code used for processing the data can be found at https://github.com/wuqui/IndVar. The dataset and R-Code used for the regression models can be found in the Supplementary Material.

AUTHOR CONTRIBUTIONS

H-JS developed the conception and design of the study; QW conducted the data retrieval and processing; HK and SF developed the conception of the statistical analyses and SF performed them; H-JS wrote the first draft of the manuscript; QW wrote the first draft of ***Data source* and *Data retrieval***; all authors contributed to manuscript revision and read and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.547531/full#supplementary-material.

REFERENCES

Anthonissen, L. (2020b). Special passives across the life span: cognitive and social mechanisms. PhD dissertion. Munich (UK): Antwerp University.

Anthonissen, L. (2020a). 'Cognition in construction grammar: connecting individual and community grammars'. *Cogn. Linguist.* 31 (2), 309-337. doi:10.1515/cog-2019-0023

Barlow, M. (2013). Individual differences and usage-based grammar. *Int. J. Corpus Linguist.* 18 (4), 443–478. doi:10.1075/ijcl.18.4.01bar

Baxter, G., and Croft, W. (2016). Modeling language change across the lifespan: individual trajectories in community change. *Lang. Var. Change* 28 (2), 129–173. doi:10.1017/s0954394516000077

Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: what corpus data do and do not reveal about the mind*. Berlin: Mouton De Gruyter.

Blumenthal-Dramé, A. (2017). "Entrenchment from a psycholinguistic and neurolinguistic perspective," in *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge*. Editor H.-J. Schmid (Washington and Boston, US: APA and de Gruyter Moulton), 129–152.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognit. Psychol.* 18 (3), 355–387. doi:10.1016/0010-0285(86)90004-6

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Appl. Linguist.* 25, 431–447. doi:10.1093/applin/25.4.431

Dąbrowska, E. (2015). "Individual differences in grammatical knowledge," in *Handbook of cognitive linguistics*. Editors E. Dąbrowska and D. Divjak (Boston, MA: Mouton de Gruyter), 650–668.

Dąbrowska, E. (2016). *"Language in the mind and in the community"*, in *Change of paradigms - new paradoxes. Recontextualizing language and linguistics*. Editors J. Daems, E. Zenner, K. Heylen, D. Speelman, and H. Cuyckens (Boston, MA: De Gruyter Mouton), 221–236.

Desagulier, G. (2014). Aneractive R script for a sociolinguistic exploration of the spoken component of the Bnc-2014. *Around the word*, Vol. 03/01/2019.

Diewald, G. (2011). Pragmaticalization (defined) as grammaticalization of discourse functions. *Linguistics* 49 (2), 365–390. doi:10.1515/ling.2011.011

Du Bois, J. W. (2014). Towards a dialogic syntax *Cogn. Linguist.* 25 (3), 359–410. doi:10.1515/cog-2014-0024

Ellis, N. C. (2017). "Chunking in language usage, learning and change: I Don't Know," in *The changing English language: psycholinguistic perspectives*. Editors M. Hundt, S. Mollin, and S. E. Pfenninger (Cambridge: Cambridge University Press), 113–147.

Fonteyn, L., and Nini, A. (2020). Individuality in syntactic variation: an investigation of the seventeenth-century gerund alternation. *Cogn. linguist.* 31 (2), 279-308. doi:10.1515/cog-2019-0040

Fressseman, D. L., and Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *Mod. Lang. J.* 92 (2), 200–213. doi:10.1111/j.1540-4781.2008.00714.x

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11 (2), 127–138. doi:10.1038/nrn2787

Geeraerts, D. (2017). "Entrenchment as onomasiological salience," in *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge*. Editor H.-J. Schmid (Boston, MA: APA and walter de Gruyter), 153–174.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: communication, context, and consequence. *Contexts accommodation: Dev. appl. sociolinguistics* 1, 1–68. doi:10.1017/cbo9780511663673.001

Giles, H., and Ogay, T. (2007). "Communication accommodation theory," in *Explaining communication: contemporary Theories and exemplars*. Editors B. B. Whaley and W. Samter (Mahwah, NJ: Lawrence Erlbaum), 293–310.

Gobet, F., Lane, P.C. R., Croker, S., Cheng, P.C.-H., Jones, G., Oliver, I., et al. (2001). Chunking mechanisms in human learning. *Trends Cognit. Sci.* 5 (6), 236–243. doi:10.1016/s1364-6613(00)01662-4

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E. (2019). *Explain me this: creativity, competition, and the partial productivity of constructions*. Princeton, NJ: Princeton University Press.

Gries, S. T., and Hilpert, M. (2010). Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *Engl. Lang. Ling.* 14 (3), 293–320. doi:10.1017/s1360674310000092

Guy, G. R. (1980). "Variation in the group and the individual: the case of final stop deletion," in *Locating language in time and space*. Editor W. Labov (New York, NY: Academic Press), 1–36.

Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London,UK: Longman.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.

Huang, Y., and Rao, R. P. N. (2011). Predictive coding. *WIREs Cogn. Sci.* 2 (5), 580–593. doi:10.1002/wcs.142

Kroczek, L. O. H., and Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Sci. Rep.* 7 (1), 17581. doi:10.1038/s41598-017-17907-9

Labov, W. (1966). *The social stratification of English in New York city*. Washington, US: Center for Applied Linguistics.

Langacker, R. W. (1987). *Foundations of cognitive grammar. Vol. I: theoretical prerequisites*. California, CA: Stanford University Press.

Mantlik, A., and Schmid, H.-J. (2018). "That-Complementizer omission in N+Be+That-Clauses – register variation or constructional change?," in *The noun phrase in English: past and present*. Editors A. Ho-Cheong Leung and W. V. d Wurff (Amsterdam/Philadelphia, PA: Benjamins), 187–222.

Nakagawa, S., and Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol. Rev.* 85 (4), 935–956. doi:10.1111/j.1469-185X.2010.00141.x

Nelson, R. (2018). How 'chunky' is language? Some estimates based on Sinclair's Idiom Principle. *Corpora* 13 (3), 431–460. doi:10.3366/cor.2018.0156

Petré, P. (2016). EEBOCorp concordancer 1.7. Available at: https://Lirias.Kuleuven.Be/Handle/123456789/436853. Leuven (Accessed February 15, 2020).

Petré, P., and Anthonissen, L. (2020). Individuality in complex systems: a constructionist approach. *cogn. linguist.* 31 (2), 185-212. doi:10.1515/cog-2019-0033

Petré, P., and Van de Velde, F. (2018). The real-time dynamics of the individual and the community in grammaticalization. *Language* 94 (4), 867–901. doi:10.1353/lan.2018.0056

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27 (2), 169–190. doi:10.1017/s0140525x04000056

Pickering, M. J., and Ferreira, V. S. (2008). Structural priming: a critical review *Psychol. Bull.* 134(3): 427–459. doi:10.1037/0033-2909.134.3.427

Schmid, H.-J. (2007). "Entrenchment, salience, and basic levels," in *The Oxford handbook of cognitive linguistics*. Editors D. Geeraerts and H. Cuyckens (Oxford: Oxford University Press), 117–138.

Schmid, H.-J. (2014). "Lexico-Grammatical patterns, pragmatic associations and discourse frequency," in *Constructions collocations patterns*. Editors T. Herbst, H.-J. Schmid, and S. Faulhaber (Berlin/New York, NY: Mouton de Gruyter), 239–293.

H.-J. Schmid (Editor) (2017a). *Entrenchment and the psychology of language learning. How we reorganize and adapt linguistic knowledge*. Boston, MA/Berlin: APA and Walter de Gruyter.

Schmid, H.-J. (2017b). "Linguistic entrenchment and its psychological foundations," in *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge*. Editors H.-J. Schmid (Boston, MA/Berlin: APA and walter de Gruyter), 435–452.

Schmid, H.-J. (2020). *The dynamics of the linguistic system. Usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.

Schmid, H.-J., and Mantlik, A. (2015). Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles. *Anglia* 133 (4), 583–623. doi:10.1515/ang-2015-0056

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of York English: was/were variation as a case study for statistical practice. *Lang. Var. Change* 24 (2), 135–178. doi:10.1017/s0954394512000129

Verhagen, V., Mos, M., Backus, A., and Schilperoord, J. (2018). Predictive language processing revealing usage-based variation. *Lang. cogn.* 10 (2), 329–373. doi:10.1017/langcog.2018.4

Walker, J. A., and Meyerhoff, M. (2013). "Studies of the community and the individual," in *Oxford handbook of sociolinguistics*. Editors R. Bayley, R. Cameron, and C. Lucas (Oxford: Oxford University Press), 175–194.

Weinreich, U., Labov, W., and Herzog, M. I. (1968). Empirical foundations for a theory of language change," in *Directions for historical linguistics. A symposium*. Editors W. P. Lehmann and Y. Malkiel (Austin, TX: University of Texas Press), 95–195.

Wolfram, W., and Beckett, D. (2000). The role of the individual and group in earlier african American English. *Am. Speech* 75, 3–33. doi:10.1215/00031283-75-1-3

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 27 January 2021
doi: 10.3389/frai.2020.523920

Avoiding Conflict: When Speaker Coordination Does Not Require Conceptual Agreement

Alexandre Kabbach[1,2]* *and Aurélie Herbelot*[2,3]*

[1]Department of Linguistics, University of Geneva, Geneva, Switzerland, [2]Center for Mind/Brain Sciences, University of Trento, Trento, Italy, [3]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

OPEN ACCESS

Edited by:
Dong Nguyen,
Alan Turing Institute,
United Kingdom

Reviewed by:
Yao-Ying Lai,
National Chengchi University, Taiwan
Steven R. Wilson,
University of Edinburgh,
United Kingdom

****Correspondence:***
Alexandre Kabbach
alexandre.kabbach@unige.ch;
Aurélie Herbelot
aurelie.herbelot@unitn.it

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *31 December 2019*
Accepted: *19 October 2020*
Published: *27 January 2021*

Citation:
Kabbach A and Herbelot A (2021)
Avoiding Conflict: When Speaker
Coordination Does Not Require
Conceptual Agreement.
Front. Artif. Intell. 3:523920.
doi: 10.3389/frai.2020.523920

In this paper we discuss the *socialization hypothesis*—the idea that speakers of the same (linguistic) community should share similar concepts given that they are exposed to similar environments and operate in highly-coordinated social contexts—and challenge the fact that it is assumed to constitute a prerequisite to successful communication. We do so using *distributional semantic models* of meaning (DSMs) which create lexical representations via latent aggregation of co-occurrence information between words and contexts. We argue that DSMs constitute particularly adequate tools for exploring the socialization hypothesis given that 1) they provide full control over the notion of background environment, formally characterized as the training corpus from which distributional information is aggregated; and 2) their geometric structure allows for exploiting alignment-based similarity metrics to measure inter-subject alignment over an entire semantic space, rather than a set of limited entries. We propose to model *coordination* between two different DSMs trained on two distinct corpora as *dimensionality selection* over a dense matrix obtained via Singular Value Decomposition This approximates an ad-hoc coordination scenario between two speakers as the attempt to align their similarity ratings on a set of word pairs. Our results underline the specific way in which linguistic information is spread across singular vectors, and highlight the need to distinguish *agreement* from mere *compatibility* in alignment-based notions of conceptual similarity. Indeed, we show that *compatibility emerges from idiosyncrasy* so that the unique and distinctive aspects of speakers' background experiences can actually facilitate—rather than impede—coordination and communication between them. We conclude that the socialization hypothesis may constitute an unnecessary prerequisite to successful communication and that, all things considered, communication is probably best formalized as the cooperative act of *avoiding conflict*, rather than maximizing agreement.

Keywords: communication, coordination, alignment, conceptual variability, distributional semantic models, similarity, socialization hypothesis

1 INTRODUCTION

Psychological approaches to semantic and conceptual knowledge rely on intertwined yet distinct notions of *concepts* and *words* (Malt et al., 2015; Malt, 2019): concepts are "the building blocks of thought" taken to be crucial to cognition at large (Margolis and Laurence, 2019), while words are "the smallest linguistic expressions conventionally associated with non-compositional meaning [. . .] which can be articulated in isolation to convey semantic content" (Gasparri and Marconi, 2019). Those psychological approaches—also referred to as *cognitivist* or *subjectivist* (Gärdenfors, 2014; Barsalou, 2017; Pelletier, 2017)—assume concepts, unlike words, to be *private* mental entities, which poses a major challenge for communication, for how could two speakers communicate if the words they utter do not refer to identical concepts? (Fodor, 1977; Pelletier, 2017).

The solution to this conundrum, we are told, lays in the inherently social nature of the lexical acquisition process (Clark, 1996; Murphy, 2002; Barsalou, 2017) for if children do acquire lexical items by matching new words to previously learned concepts (e.g., Bloom, 2000) they do not do so randomly: they learn through socialization which concepts go with which words, so that the internal mental representations associated with words are shaped by many years of interactions with other speakers of the same (linguistic) community. As a result, speakers of the same community relate words to very *similar* concepts (Murphy, 2002, p. 391). The *socialization hypothesis*—as we propose to name it—therefore postulates that speakers of the same community *should* share similar concepts given that they are exposed to similar environments and operate in highly-coordinated social contexts (see **Section 2.1**).

Yet, conceptual similarity remains *hard* to validate experimentally, and is more often than desired a matter of seeing the glass as half full: speakers never significantly disagree on their judgments of similarity, but never totally agree either (see **Section 2.2**). Meanwhile, recent work in cognitive science has attempted to come to term with the idea that concepts may vary widely across individuals, some even suggesting that it may not necessarily represent an obstacle to communication, as what matters ultimately is that speakers coordinate *during* conversation and *align* their conceptual representations on aspects relevant to the situation under discussion (see **Section 2.3**).

Yet again, this notion of *alignment* remains dubious as it is often relaxed to mere *similarity* or *sufficient overlap*. But what does it *mean* for two concepts to be similar? And how much similarity is *enough* for successful communication? In fact, alignment-based similarity appears more often than not to be a matter of overall *compatibility* rather than strict *agreement*: being highly tolerant to variability, it can potentially settle for minimal overlap so that speakers holding marginally identical conceptual representations can still be assumed to understand one another. But if *anything goes*, then this notion of similarity becomes rather devoid of content and pretty much useless for assessing the pertinence of the socialization hypothesis.

As always, the devil is in the details. For indeed the socialization hypothesis focuses on conceptual *spaces* and as such pertains to the *whole structure* rather than the *superficial parts*. After all, the notion of conceptual variability considered so far remains superficial in as much as it is only observed through the lens of limited behavioral response patterns in humans. And since *superficial variability does not preclude latent structural similarity*, conceptual spaces could still very well be aligned despite the apparent variability, provided the adequate characterization of alignment (see **Section 2.4**). Additional methodological challenges still remain in order to validate the socialization hypothesis, for 1) it is never possible to gain full access over speakers' background experiences which presumably condition the formation of their respective conceptual spaces; and 2) it is in practice never possible to test human subjects on their entire lexicons, let alone conceptual spaces, in order to guarantee the robustness of the observed experimental results.

To overcome parts of those methodological challenges, we propose in this work to rely on distributional semantic models of lexical meaning (DSMs) which create vector representations for words via latent aggregation of co-occurrences between words and contexts (see **Section 3**). We argue that those models prove particularly suited for assessing the validity of the socialization hypothesis, given that 1) they provide full control over speakers' background experiences, formalized experimentally as the training corpus from which distributional information is aggregated; 2) their geometric structure allows for exploiting alignment-based similarity metrics to measure inter-subject alignment, and do so over an entire semantic space rather than a set of limited entries, thereby overcoming the experimental shortcomings of testing on human subjects; and 3) their overall generation pipeline parallels humans' conceptual processing in a cognitively plausible fashion.

Following the core assumptions underpinning the socialization hypothesis stated above, we propose to distinguish within our model *background experience* from *active coordination*. On the one hand, we control for background experience by varying the data fed to the DSM. On the other hand, we implement *active coordination* by modifying the standard DSM pipeline, which normally includes a dimensionality reduction step involving the top singular vectors of a Singular Value Decomposition (SVD). Specifically, we replace the variance-preservation bias by an explicit coordination bias, sampling the set of d singular vectors which maximize the correlation with a particular similarity dataset (see **Section 4.1**). Thereby, we approximate an ad-hoc coordination scenario between two speakers as the attempt to align their similarity ratings on a set of word pairs. We then propose to quantify structural alignment between two DSMs as the residual error between their two matrices, measured after having put their elements in correspondence with one-another (see **Section 4.2**).

Using the above methodology, the paper makes three contributions. First, we show that *no variance-preservation bias means better superficial alignment*. Indeed, we show that replacing the variance-preservation bias by an explicit sampling bias leads to near-systematic improvements on various lexical similarity datasets. We show in addition that this result is fundamentally grounded in the fact that *different dimensions in the SVD encode different semantic phenomena*, so that DSMs can actually capture a collection of possible meaning spaces from the same set of data, rather than a single one (see **Section 5.1**).

Second, we show that *better superficial alignment does not mean better structural alignment*. Although alignment is arguably a complex and multifaceted process, we show that, when considered from the point of view of our specific characterization, the systematicity of the relation between superficial and structural alignment does not hold (see **Section 5.2**).

Third, we show that conceptual spaces generated from different background experiences can be aligned in different ways, and that the aforementioned considerations over *alignment* and *compatibility* extend from conceptual *representations* to conceptual *spaces*. Indeed, we show that DSMs can be aligned by sampling pairs of singular vectors which highly correlate with one another, but also very often by sampling singular vectors that do not correlate but nonetheless increase the structural similarity between the two modeled conceptual spaces (see **Section 5.3**). A deeper investigation of this effect suggests that *compatibility emerges from idiosyncrasy*, so that the unique and distinctive aspects of speakers' background experiences can actually facilitate—rather than impede—coordination and communication between them (see **Section 6**).

We conclude that the socialization hypothesis may constitute an unnecessary prerequisite to successful communication and that, all things considered, communication is probably best formalized as the cooperative act of *avoiding conflict*, rather than maximizing agreement.

2 CONCEPTUAL VARIABILITY AND THE SOCIALIZATION HYPOTHESIS

2.1 The Socialization Hypothesis: Review and Overview

The primary observation underpinning the socialization hypothesis is that conceptual acquisition precedes lexical acquisition, so that children first acquire concepts before learning to map them to corresponding lexical labels (Clark, 1983; Mervis, 1987; Merriman et al., 1991; Bloom, 2000). The key idea behind the hypothesis is then to consider that the acquisition of this conceptual-to-lexical mapping is not random but rather heavily constrained, in that it takes place in a highly coordinated social context, so that speakers of the same community end up assigning similar concepts to the same words. Phrased along those lines, the hypothesis can be found in (Murphy, 2002, p. 391):

> [. . .] people do not associate any old concept to a word. Instead, they learn through socialization which concepts go with which words. So, as a child, you learned that dog refers to a certain kind of animal. If you first developed the hypothesis that dog refers to any four-legged mammal, you would soon find yourself miscommunicating with people. Theywould not understand you when you referred to a sheep as dog, and you would not understand them when they said that all dogs bark, and so on. Thus, there is a social process of converging on meaning that is an important (and neglected) aspect of language [. . .]

However, the socialization hypothesis extends beyond the conceptual-to-lexical mapping itself: since human beings should have similar cognitive systems and evolve in similar environments overall, they should end up sharing similar *conceptual spaces* (Barsalou, 2017, p. 15):

> [. . .] different individuals have similar bodies, brains, and cognitive systems; they live in similar physical environments; they operate in highly-coordinated social contexts. As a result, different individuals acquire similar distributed networks for a given concept over the course of development. Within a particular social group or culture, different individuals' networks are likely to be highly similar, given similar coordinated experiences with many shared exemplars. Even across different cultures, these networks are likely to be highly similar, given that all humans have similar bodies, brains, and cognitive systems, operating in similar physical and social environments.

In both Murphy's and Barsalou's formulations of the hypothesis we find the idea that there are both individual and collective—cognitive and social—processes at play in both conceptual and lexical acquisition, as well as linguistic communication as a whole. The underlying idea is that people *cooperate* with one another when they use language (Austin, 1962; Grice, 1975) and perform what Clark (1996) has called *joint actions* on top of individual actions, so that they coordinate with one another in order to converge to some *common ground* (Clark, 1992; Clark, 1996). This notion of common ground (see also Stalnaker, 2002; Stalnaker, 2014) encompasses notions of *common knowledge* (Lewis, 1969), *mutual knowledge* or *belief* (Schiffer, 1972) and *joint knowledge* (McCarthy and Lifschitz, 1989) and covers whatever knowledge or beliefs speakers of the same (linguistic and/or cultural) community may share. It also includes what Gärdenfors (2014) refers to as *third-order intersubjectivity*: not only what I know, but also what I assume you know and what I assume you know that I know. Overall, the general idea put forth by Clark (1996) is that the more time people spend together, the larger their common ground; an idea which we can re-interpret in light of the socialization hypothesis as *shared experiences entail shared conceptual spaces*.

But coordination is also a process which takes place at the lexical level so that speakers can settle for a particular word meaning, a phenomenon that Clark (1992) has called *entrainment*.[1] As such, and in as much as the socialization

[1] For earlier work on lexical coordination focusing on reference, see (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996).

hypothesis can be said to presuppose meaning to derive from *convention*, one can trace its foundational considerations to Plato's *Cratylus* (Cooper, 1997) and its discussion on the essence of meaning. According to Rescorla (2019), there is now a wide consensus in philosophy to stand with Hermogenes against Cratylus in considering that language at large is conventional, in that the association between a word and its referent is arbitrary and driven by convention rather than intrinsic to the nature of words. Conventional views of meaning have given rise to a very rich literature since the *signaling games* of Lewis (1969) which have proposed a formal characterization of the phenomenon of semantic convergence, grounded in Gricean pragmatics and the idea that meaning emerges from active coordination between speakers' communicative intentions and hearers' expectations (Grice, 1969).

Conventional views of meaning do not preclude however the semantics of a word to vary across time, or even across utterances. Cruse for instance, has argued that the meaning of a word changed to some extent at each of its occurrences—what he has called *context modulation* (Cruse, 1986, p. 52). Barker (2002) further observed that utterances could shift the meaning of a predicate, and those considerations have led several researchers to propose the idea of the existence of a *core* meaning for each word sense, core meaning potentially pragmatically modulated at each utterance (Lasersohn, 1999; Recanati, 2004; Wilson and Carston, 2007). Such considerations extend to concepts at large and the question of whether or not they have *cores* themselves (see Barsalou, 2017, for an overview). Indeed, several proposals have been made to argue against the notion of conceptual core and for the idea that concepts are, in part of in full, context-dependent (Evans, 2009; Connell and Lynott, 2014; Casasanto and Lupyan, 2015). This argument is partly supported by empirical evidence showing that not all conceptual information, even what could be considered central one, is automatically activated across context (Kiefer et al., 2012; Gawronski and Cesario, 2013; Lebois et al., 2015).

However, and despite the above consideration over conceptual variability, the socialization hypothesis remains grounded in the idea that identicity of concepts across speakers is not necessary for successful communication: sufficient conceptual *overlap* or *similarity* suffice. This idea can be found as early as (Humboldt, 1836/1988, p.152), when stating that:

> Men do not understand one another [. . .] by mutually occasioning one another to produce exactly and completely the same concept; they do it by touching in one another the same link in the chain of their sensory ideas and internal conceptualizations, by striking the same note on their mental instrument, whereupon matching but not identical concepts are engendered in each.

Relaxing the constraint over conceptual identicity across subjects remains nonetheless problematic, for it pushes the burden of proof over to the notion of similarity: what does it mean for two concepts to be *similar*? And how much similarity is *enough* for successful communication? (see, e.g., Connell and Lynott, 2014, p. 400). As we will see in the following section, unequivocally aligning similarity judgments is difficult to achieve across human subjects, and the proper characterization of similarity remains both a theoretical and an experimental challenge, so that the question of whether or not two speakers hold similar conceptual spaces is sometimes left to seeing the glass as half full.

2.2 Conceptual Similarity: An Experimental Challenge

What does it *mean* to *hold* a concept? As a first approximation, Murphy (2002) proposes to assimilate conceptual knowledge to lexical knowledge, although it has been convincingly argued that words do not begin to capture the richness of their underlying conceptual representations (Landau et al., 2010; Wolff and Malt, 2010; Gleitman and Papafragou, 2012). Marconi (1997) proposes to further distinguish within lexical knowledge the notion of *inferential* competence—the ability to *name* objects—from the notion of *referential* competence—the ability to *refer* to objects. This distinction is supported by empirical evidence from neuroscience showing that certain brain pathologies may affect one competence while leaving the other intact (Warrington, 1975; Heilman et al., 1976; Kemmerer et al., 2012; Pandey and Heilman, 2014). Marconi (1997) takes it for granted that lexical competence may vary widely across speakers of the same language, for language reflects what Putnam (1975) has called the *division of linguistic labor* which derives from the division of *non*-linguistic labor. That is, knowledge effects entailed by differences in expertise on a given domain may translate as differences in lexical knowledge across speakers. Yet, Marconi still assumes that certain parts of the lexicon will remain preserved from the interference of specialized knowledge, so that lexical competence for a certain number of words can be considered reasonably identical across speakers. He takes the word *spoon* to be one such example (Marconi, 1997, p. 57), and yet Labov (1973) showed in his seminal work on the semantics of tableware items that denotation for words such as *mug*, *cup*, *bowl* and *vase* could vary widely across individuals when modifying objects properties such as *width*, *depth*, *content* or even *presence or absence of a handle*. Labov's study illustrates what has since been confirmed over and over experimentally, and what Pelletier summarizes as the fact that "different subjects give individually different results on the many tasks about meaning that have been administered over the decades in cognitive psychology" (Pelletier, 2017, p. 74). Indeed, psychological experiments on lexical similarity—which typically ask subjects to grade lists of word pairs on a ten-point scale, or triangular arrays of words by choosing among a pair of word the most similar to a referent word (Hutchinson and Lockhead, 1977)—exhibit mixed levels of agreement across subjects: from 0.44 to 0.63 on word pairs and from 0.45 to 0.66 on triangular arrays depending on the categories being tested (e.g., fruits or birds; see Hutchinson and Lockhead, 1977, p. 667).

Those results could be considered artifactual of experimental setups artificially decontextualizing lexical items by presenting them in isolation and without sentential context—potentially ignoring thereby the effect of *context modulation* (see **Section 2.1**). And indeed Anderson and Ortony (1975) confirmed experimentally that subjects modulate the meaning of a word at least based on the sentence in which it occurs. Murphy and Andrew (1993) even showed that subjects could change their judgments over synonyms and antonyms depending on the presented word pairs. Nonetheless, even experiments which do try to evaluate human similarity judgments in heavily constrained contextual setups exhibit non-trivial inter-speaker variability: in their study comparing lexical expectations across individuals, Federmeier and Kutas (1999) presented subjects with clozed sentence pairs such as *They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of* … and three target words comprising an expected exemplar (e.g., *palms* for the above example), an unexpected exemplar of the same category (e.g., *pines*) and an unexpected exemplar of a different category. Expectations regarding missing words were first evaluated as clozed probabilities computed by asking a set of subjects to select the best target candidate given the presented context, but only averaged at 0.74 while ranging from 0.17 to 1 depending on tested items. Other lexical substitution experiments performed on humans exhibit similarly low agreement levels across subjects: 0.28 for McCarthy and Navigli (2009) and as low as 0.19 and 0.16 for Kremer et al. (2014) and Sinha and Mihalcea (2014).

Could such relatively moderate levels of agreement constitute mere byproducts of the unreliability of introspective judgment? The question is not quite settled: Federmeier and Kutas (1999) did attempt to analyze the distribution of N400 across subjects—a negative-going potential peaking around 400ms after stimulus onset which often indicates semantic anomaly or an unexpected event. Yet, and although they did find slight differences in N400 patterns across subjects, they blamed the intrinsic variation of brainwaves across individuals and did not investigate further given the relatively small size~(6) of their sample of participants.

Of course, one could also say that lab experiments operatenecessary methodological approximations which lead to unrealistic language usage setups that do not, all things considered, invalidate the socialization hypothesis: communication is not a clozed test, let alone a lexical similarity task. Lexical variability at the word level, even if attested experimentally, does not preclude conceptual similarity to be validated when language takes places in a realistic, articulated, and coordinated communication setting. Words are seldom if ever used in isolation to refer to their underlying conceptual representations, and vice versa. Yet, inter-speaker variations in concept-to-word mappings led to very concrete problems when attempting to design verb-mediated computer interfaces in the 1990s: Furnas et al. (1987) for instance showed that agreement on (computer-) function-to-word mapping ranged from 0.07 to 0.18, and agreement on word-to-function mapping remained at 0.15 (see also Brennan, 1998). In other words, subjects barely used the same word to refer to the same function/concept, or thought of identical functions/concepts when using the same word, rendering verb-mediated computer interfaces practically unusable.

The notion of (conceptual and/or semantic) *similarity* itself is a challenge: it varies with experience, knowledge, expertise or even (linguistic) context (see Medin et al., 1993; Goldstone and Son, 2012, for an overview). Its theoretical foundations are somehow shaky, for A is always similar to B *with respect to something* (Goodman, 1972). Therefore, it pushes yet again the burden of proof over to modeling considerations on the notion of *context*, especially as similarity judgments remain sensitive to *tasks* (Murphy and Medin, 1985) and *instructions* (Melara et al., 1992).

We could still acknowledge the ubiquity of conceptual variability across speakers but postulate nonetheless that the notion of similarity should pertain to a more stable or invariant part of the conceptual structure. Prototypes (Rosch, 1973; Rosch, 1975; Rosch, 1978) could form such a proposal for conceptual invariance, and yet they also prove sensitive to context (Roth and Shoben, 1983). Moreover, the stability of prototypical structure across subjects may not be as high as originally demonstrated, as Barsalou (1987) showed on a large-scale replication study that inter-subject agreements on prototypes ranged between 0.45 and 0.50, significantly below the original 0.90 reported by Rosch (1975).

Assessing conceptual similarity experimentally is subject to many interfering parameters. One of them, as we previously mentioned, is *knowledge* (Goldstone and Son, 2012). Several proposals have been made to bypass knowledge interference, one of them being to experiment on dummy or artificial concepts which specifically require no previous knowledge from tested subjects (Murphy, 2002, p. 141). Yet again, similarity judgments based on artifact categories have proven unreliable as artifact categories are unstable and depend on the categorization task at hand (Sloman and Malt, 2003; Malt and Sloman, 2007).

In short, conceptual similarity remains *hard* to validate experimentally, and is more often than desired a matter of seeing the glass as half full: speakers never significantly disagree on their similarity judgments, but they never totally agree either. The pervasiveness of conceptual variability has gradually worked its way through cognitive science, and much recent work now take for granted that conceptual representations can never be assumed to be fully identical across speakers, given that they are essentially grounded in different background experiences (e.g., Connell and Lynott, 2014, p. 400). For some, it should be relatively easy to come to term with the idea that speakers hold rather different concepts, given how often linguistic communication actually requires clarification (Yee and Thompson-Schill, 2016, p. 1024). For many, however, this still does not necessarily represent an obstacle to successful communication, as what matters ultimately is that speakers are able to coordinate *during* conversation to align their conceptual representations on aspects relevant to the situation under discussion (e.g., Pickering and Garrod, 2006; Connell and Lynott, 2014). We now turn to a historical overview of those approaches and to what their formal characterizations entail.

2.3 From Coordination to Alignment

As we have previously detailed in **Section 2.1**, linguistic communication requires *cooperation* and *coordination* between interlocutors in that it notably involves speakers doing things with words while trying to have their addressees recognize their intentions (Clark, 1992, p. xii). As Clark (1996) emphasized, there is more to language than just a speaker speaking and a listener listening, thus linguistic communication cannot be reduced to mere signal processing. Several research have therefore since proposed to approach (linguistic) communication as *alignment of information states* rather than *information transfer* (e.g., Pickering and Garrod, 2004; Pickering and Garrod, 2006; Garrod and Pickering, 2009; Pickering and Garrod, 2013; Wachsmuth et al., 2013). Speakers and addresses, they argue, are not rigid entities but interactive agents, constantly negotiating meaning during conversation while relying on dynamic and perpetually evolving conceptual representations. Coordination, then, should be understood as the process by which interlocutors converge to similar if not identical mental representations during conversation, a process referred to as *alignment* (Pickering and Garrod, 2004, p. 172).

Interactive-alignment-based models of linguistic communication such as (Pickering and Garrod, 2004; Pickering and Garrod, 2006) distinguish what they call *situation models* from *linguistic representations* and *general knowledge*. A situation model is defined as a multi-dimensional representation of the situation under discussion—encoding space, time, causality, intentionality and reference to main individuals under discussion (Zwaan and Radvansky, 1998)—and is assumed to capture what people are "thinking about" during conversation. The embodied (and embedded) approach to cognitive science operates a similar distinction between *representations* and *concepts*. A *representation* refers to a "specific, situated, contextual instantiation of one or more concepts necessary for the current task", while a concept refers to "a general, aggregated, canonical (i.e., contextfree) aspect of experience that has the potential to form the basis of an offline representation" (Connell and Lynott, 2014, pp. 391–392). The distinction between (online) representations and (offline) concepts allows the aformentioned approaches to overcome the challenge posed by conceptual variability to communication: offline concepts may differ widely across interlocutors, successful communication remains possible provided that online representations—or situation models—can be aligned (see, e.g., (Pickering and Garrod, 2006, p. 204) or (Connell and Lynott, 2014, p. 400)).

The way in which those approaches accommodate conceptual variability remains nonetheless quite relative, all things considered. First of all, because they assume coordination to play a key role in the socialization hypothesis itself. Indeed, they do not expect concepts and representations to develop in isolation, but rather to mutually influence one another: online representations or situation models are expected to draw upon both linguistic and general (conceptual) knowledge (Connell and Lynott, 2014, pp. 391–392) while, in return, *online perception affects offline representation* (see Principle 1 in Connell and Lynott, 2014, p. 393). Moreover, they assume that alignment at one level of representation will enable or improve alignment at other levels (Pickering and Garrod, 2004, p. 172) so that speakers are expected to align their general knowledge—and the underlying concepts—alongside their situation models throughout coordination (Pickering and Garrod, 2006, p. 215). Consequently, coordination is considered to act as a catalyzer of conceptual similarity: it is not only that speakers of the same community will be better able to coordinate thanks to the similarity of their conceptual spaces—itself deriving from the similarity of their background experiences—it is also that repeated coordination between them will in turn increase their overall conceptual similarity, ultimately leading to a virtuous circle of mutual understanding across speakers of the same community.[2]

Second of all, and more importantly, the tolerance of the aforementioned approaches to conceptual variability remains all relative in that they still consider similarity between background experiences to constitute a prerequisite to successful alignment, coordination and therefore communication. As Garrod and Pickering (2009) point out, "alignment is typically achieved [...] because people start off at a very good point. They communicate with other people who are largely similar to themselves, both because they process language in similar ways and because they share much relevant background knowledge" (see p. 294). As such, they rest upon a strong interpretation of the socialization hypothesis, where it should *not* be possible for any two speakers to coordinate and therefore successfully communicate if their respective conceptual spaces remain grounded in fundamentally different background experiences. In fact, the socialization hypothesis still remains a prerequisite to successful communication.

Those considerations invariably lead us to question how strictly we should understand the notion of alignment so far defined to entail *identicity* of conceptual representations. After all, given that online representations are expected to draw upon both linguistic and offline conceptual knowledge, alignment should always be partial at best (Pickering and Garrod, 2006, p. 215). But the interactive-alignment-based models remain heavily grounded in the Shannon–Weaver code model of communication (Shannon and Weaver, 1949) and as such they still often explicitely consider *identicity of messages* between interlocutors to define communication success (see, e.g., Pickering and Garrod, 2013, p. 329). Yet again, this identicity constraint is often relaxed to mere similarity or sufficient overlap (e.g., Connell and Lynott, 2014, p. 400) and successful communication under conceptual misalignment is then considered possible, but only in as much as misalignment pertains to aspects of conceptual knowledge that are irrelevant to the conversation at hand (Pickering and Garrod, 2006, p. 215). The following example, adapted from (Connell and Lynott, 2014, p. 401) illustrates how, in fact, alignment may

[2] The role of coordination in the socialization hypothesis is explicit in Barsalou's characterization introduced in **Section 2.1**.

not always equate *agreement* but sometimes mere *compatibility* between conceptual representations:

> [. . .] imagine your lifetime experience of dogs has been entirely of the small, handbag-dog variety, and that you are unaware that dogs come in any form larger than a chihuahua. You then meet someone who has only ever experienced large working dogs and is unaware that dogs come in any form smaller than a German shepherd. An exchange such as "Do you like dogs?" "Yes, we have one at home," "Same here, we just got one last week from the shelter," is perfectly effective communication where each party understands the other, even though each individual is representing quite a different dog in both canonical (i.e., liking dogs in general) and specific (i.e., my pet dog at home) forms [. . .]

The question, then, pertains to the prevalence of compatibility: should it be considered the norm rather than the exception? And how far does it extend? For if indeed the notion of similarity so far considered actually tolerates extreme ranges of variability and negligible overlap between conceptual representations, then it becomes rather devoid of content. Even more so if, as we later show in **Section 6**, compatibility emerges from idiosyncrasies in speakers' background experiences, so that alignment can be satisfied even with conceptual representations grounded in fundamentally different background experiences. And the socialization hypothesis then becomes unnecessary, if not inoperative. Before we turn to a more formal investigation of the questions at hand, let us detail several remaining theoretical and methodological challenges.

2.4 Remaining Obstacles to the Formal Characterization of the Socialization Hypothesis

As we have previously emphasized in **Section 2.1**, the socialization hypothesis is first and foremost a hypothesis about conceptual *spaces*. As such, it rests upon a very important property of human cognition at large, namely, that the conceptual space *has structure* (Gärdenfors, 2004; Gärdenfors, 2014).

This particular emphasis on the structure of the conceptual space stresses the need to operate a distinction between latent *structure* and *surface* form, especially when it comes to alignment. This distinction is all the more important that Wachsmuth et al. (2013) underlined that the two do not necessarily go hand-in-hand, for, first, *superficial alignment does not necessarily guarantee structural alignment* (see p. 5). In the particular case of conceptual similarity that concerns us here, this notion of *surface form* can be understood as the behavioral response subjects typically exhibit on various cognitive tasks—such as lexical similarity judgments—the only type of empirical evidence actually accessible to us in practice, for conceptual representations within subjectivist or cognitivist approaches remain mere theoretical constructs. Yet, the problem is, as it has been long argued, that behavioral correlates between subjects on such tasks do not guarantee identicity of concepts (see, e.g., (Davidson, 1984, p. 163), or (Pelletier, 2017, p. 52)). Indeed, Gentner (1988), for instance, showed that adults and children below 8 years old respond differently to the question "how is a cloud like a sponge?": children, unlike adults, are more inclined to favor the attributional interpretation that "they are both soft and fluffy" over the relational one that "they can both hold water and give it off later". Such differences in response patterns typically exemplify discrepancies across subjects' underlying concepts of CLOUD and SPONGE, and across their relationships to other concepts such as WATER or even FLUFFINESS. Those apparent discrepancies, however, do not preclude mutual agreement on their respective judgments of similarity with respect to CLOUD and SPONGE.[3]

Conversely, Wachsmuth et al. (2013) argued that *superficial variability does not necessarily imply structural misalignment* (*ibid.*). Here again, one must bear in mind that the socialization hypothesis pertains to a *whole* that is more than just the *sum of its parts*. Yet, due to the practical limitations of experimenting on human subjects, the type of conceptual variability reported in **Section 2.2** is almost systematically aggregated on a (very) limited set of entries that may not be representative of the conceptual space *as a whole*. Therefore, it is perfectly possible that such empirical evidence does not actually call into question the socialization hypothesis, for it may not actually prevent a characterization of *overall* similarity between conceptual spaces. Even more so if we are to take into account the *division of linguistic labor* previously detailed in **Section 2.2**, which suggests that variations across speakers' conceptual representations may be unevenly distributed across the entire conceptual space, and that high local variability is actually to be expected. Thus, in addition to developing experimental protocols that allow for testing conceptual similarity across the *entirety* of the conceptual space, it appears necessary to develop measures of conceptual similarity that quantify the overall structural similarity between any two spaces, while potentially tolerating high degrees of local and superficial variability.

To overcome parts of the aforementioned challenges, we propose to resort to distributional semantic models of lexical meaning. Indeed, we argue that those models prove particularly suited for the modeling task at hand, given that 1) they provide full control over speakers' background experiences; 2) their geometric structure allows for defining two distinct notions of similarity: a) at the superficial level, between any two elements, through the notion of cosine *similarity* which models humans behavioral response to lexical similarity tasks; and b) at the structural level, between any two distributional models, through the notion of transformational alignment which makes it possible to quantify similarity over entire spaces, rather than a set of limited entries; and 3) their overall generation pipeline

[3] In case readers were to wonder whether her experimental protocol were not forcing artificial similarity judgments upon subjects, note that Gentner (1988) specifically mentions cases were children explicitly reject metaphorical interpretations for concepts they do not consider to be similar.

parallels that of human processing and conceptual formation in a cognitively plausible way. We now turn to their formal introduction.

3 DISTRIBUTIONAL SEMANTIC MODELS

3.1 Definition

Distributional Semantic Models (DSMs; Turney and Pantel, 2010; Clark, 2012; Erk, 2012; Lenci, 2018) can be formalized as tuples $< T, C, F, S >$, meaning that a set of targets T is represented in terms of a function F of the frequency of co-occurrence of its elements with a set of contexts C. S is then a measure defined over $T \times T$ that yields results interpreted as similarity judgments. DSMs have been shown to successfully account for a number of linguistic phenomena, both at the word and sentence level (see Lenci, 2018, for an overview). Their success, however, is dependent on the exact shape of the model, in particular its architecture and hyperparameters, and the fine-tuning of each of the components has been widely explored in the literature (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Baroni et al., 2014; Kiela and Clark, 2014; Lapesa and Evert, 2014; Levy et al., 2015).

DSMs come in two notable variants: *count-based* models such as those originally used for Latent Semantic Analysis (Landauer and Dumais, 1997) and *prediction-based* models which create dense representations for words by learning to predict target words and/or context words using neural networks (e.g., Collobert and Weston, 2008; Mikolov et al., 2013a; Mikolov et al., 2013c). Although Baroni et al. (2014) originally argued that prediction-based DSMs outperform their count-based counterparts, Levy et al. (2015) and Mandera et al. (2017) have since shown that both count and predict models could perform equally well provided specific modeling adjustments and hyperparameters tuning, especially as Levy and Goldberg (2014) showed that certain implementations of prediction-based models are actually equivalent to count-based ones in that they actually perform implicit matrix factorization of the PMI weighed word-context matrix. Despite all considerations, count-based models remain the more direct implementation of the distributional hypothesis of Harris (1954) and are still considered solid options for meaning representation, especially because of the increasing necessity to have transparent and explainable models.

In a traditional count-based model distributional representations of words are computed by aggregating co-occurrence counts of context words found on both sides of a target within a specified range called the *window size*. A given entry of the raw count matrix, corresponding to the row index of a target word w and the column index of a context word c is then weighted using Positive Pointwise Mutual Information (PPMI):

$$\mathrm{PPMI} = \max(\mathrm{PMI}(\mathrm{w}, \mathrm{c}),\ 0) \tag{1}$$

where the PMI for w and c is given by:

$$\mathrm{PMI}(w, c) = \log \frac{P(w, c)}{P(w) \cdot P(c)} \tag{2}$$

In order to reduce the dimensionality of the $T \times C$ matrix and to capture higher order co-occurrences that are latent in the data, the sparse PPMI matrix of word vector representations W is then converted to a dense matrix using Singular Value Decomposition ~(SVD):

$$W = U \cdot \Sigma \cdot V^{\top} \tag{3}$$

where U is the matrix of (left) singular vectors, Σ is the matrix of singular values, and V is the matrix of (right) singular vectors. W is then reduced to a low-dimensional matrix W_d by selecting the top d singular vectors ranked in decreasing order of singular values:

$$W_d = U_d \cdot \Sigma_d^{\alpha} \tag{4}$$

where the exponent $\alpha \in [0, 1]$ is a hyperparameter which has been shown to positively impact performances on some specific semantic tasks (Caron, 2001; Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Levy et al., 2015).[4]

The usual motivation behind dimensionality reduction is to drop factors that account for little variability in the original weighted PPMI matrix. In the particular case of SVD described above, the reduced matrix W_d is often referred to as the *best rank-d approximation* (e.g., Martin and Berry, 2007, p. 41). The choice of the first d dimensions therefore relies on a variance-preserving assumption: as the obtained W_d matrix is the one that best approximates, among matrices of rank d, the original PPMI matrix, it should also be the one that better represents the desired semantic space. Yet, while the hyperparameters' space has been widely explored in the literature, this assumption has hardly ever been questioned. Interestingly, we show in the following section that the preservation of the total variance in the original matrix is marginal at best, casting doubts on the original motivation behind this variance-preservation bias. As we will later show in **Section 4.1**, calling into question the variance-preservation bias proves determinant in investigating the socialization hypothesis, in that it concretely allows us to model coordination and conceptual alignment within the distributional semantics framework with only marginal modifications to the traditional DSM generation pipeline. Indeed, we show in **Section 5.1** that it is actually possible for DSMs to capture different kinds of semantics relations from the same corpus, so that rather than generating a *single* meaning space from the PPMI matrix, a *collection* of possible meaning spaces could coexist within the same set of data. Coordination then becomes the process of *dimensionality sampling*, that is, the process of reducing the SVD matrix by selecting the set of singular vectors that best satisfy the coordination constraints under consideration, rather than those that best preserve the variance.

3.2 The Variance-Preservation Bias

Bullinaria and Levy (2012) originally questioned the importance of the top singular vectors in the SVD matrix and suggested removing the first 100 dimensions, claiming that the highest variance components were influenced by aspects that turned out to be irrelevant to lexical semantics. Their observation remained

[4] We further discuss the influence of the α parameter in **Section 5.1**.

TABLE 1 | Percentage of total energy preserved with d = 10 000 and d = 300 top dimensions for DSMs trained on various corpora described in **Table 2**.

	d = 10 000	**d = 300**
WIKI07	66%	11%
OANC	72%	11%
WIKI2	58%	10%
ACL	62%	13%
WIKI4	52%	9%
BNC	59%	10%
WIKI	39%	9%

All models are PPMI-weighted count-based DSMs generated with a window of 2.

nonetheless largely ignored in the literature, and it is only very recently that research formally questioned the process of dimensionality selection in DSMs (Mu and Viswanath, 2018; Raunak et al., 2019) ultimately bringing further supporting empirical evidence to the original claim of Bullinaria and Levy (2012).

The process of dimensionality selection can be motivated by slightly different considerations: 1) creating compact and computationally efficient vector representations, which can even lead to significant performance improvement (Landauer and Dumais, 1997; Bullinaria and Levy, 2012); 2) reducing some undesirable geometrical effect in the original vector space (Grefenstette, 1994, p. 102); or even 3) mitigating the noise intrinsically present in partial data and increasing the robustness of the model (Deerwester et al., 1990). Regardless of the underpinning motivation, the dimensionality reduction process considered here remains a *lossy* process, where part of the data may be deliberately discarded following specific modeling considerations. In that sense it is to be distinguished from *rebasing* and potentially *lossless* methods which may be able to align the dimensionality of the reduced space to the original data matrix rank. An example of such approaches is *multidimensional scaling* (MDS; Shepard, 1962a; Shepard, 1962b) where similarity ratings on sets of word pairs are first collected among human subjects, before attempting to account for the entirety of the collected data via a few potentially meaningful latent dimensions in order to further explore the notion of similarity under study (Heider and Olivier, 1972; Ross and Murphy, 1999).[5]

The *best rank-d SVD approximation* that interests us here, however, is historically grounded in methodological considerations coming from image processing and more specifically image compression (e.g., Andrews and Patterson, 1976a; Andrews and Patterson, 1976b). Given an image represented as a matrix of pixels, the frequent correlation between nearby pixels in images will allow for the creation of low-dimension representations with only a few singular vectors accounting for most of the variance in the original data (Strang, 2016, p. 365). Variance-preservation is quantified via the notion of matrix *energy* (E), formally defined as the square of the Frobenius norm of the data matrix and also equal to the sum of the squared singular values of the data matrix SVD (see **Eq. 5**).

$$E_W = \left|\left|W\right|\right|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left|w_{i,j}\right|^2 = \sum_{i=1}^{\min\{m,n\}}\lambda_i^2, with \begin{cases} W \in \mathbb{R}^{m\times n} \\ W = U \cdot \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{\min\{m,n\}} \end{bmatrix} \cdot V^\top \end{cases} \quad (5)$$

A traditional *rule of thumb* for SVD dimensionality selection in image processing is to try and retain about 90% of the original energy (Leskovec et al., 2014, p. 424). Yet, as we can see in **Table 1**, this is far from being the case when selecting the top 300 dimensions of the SVD on a standard PPMI-weighted count-based DSM model, as the preserved energy remains systematically below ∼ 15%. Moreover, results on d = 10,000 suggest that the aforementioned rule of thumb is difficult to apply as-is to DSMs as it leads to high-dimensional and therefore computationally inefficient models.

This issue, however, is barely mentioned in the literature: Bullinaria and Levy (2007) explain that dimensionality reduction is performed with minimal loss defined using the standard Frobenius norm, but do not quantify it (see p. 897). Earlier work using SVD for Latent Semantic Analysis state that many of the latent singular values remain small and can therefore be ignored (Deerwester et al., 1990, p. 395). But this observation is misleading: as we can see in **Figure 1**, the distribution of singular values follows a highly-skewed Zipfian curve, so that the latent components may indeed quickly appear very small in comparison to the top components. However, the tail of the distribution remains quite *long*, especially as **Table 1** suggests the matrix rank to be significantly higher than 10,000. The cumulative effect of the tail's length can therefore be so that retaining only a few top components, even if those correspond to significantly higher singular values, may prove to account for only a tiny portion of the total energy. Be that as it may, the most frequent observation supporting the choice of a limited number of top components in the SVD remains that models simply "work" as-is, and the double benefit of having both computationally efficient and effective models frees authors from having to investigate further the consistency of their modeling choices (e.g., Lund and Burgess, 1996).

3.3 Cognitive Plausibility of DSMs

Determining whether DSMs constitute cognitively plausible models first requires asking what DSMs are supposed to be models *of*. And yet the answer to that question appears to be far from consensual: Sahlgren (2008), for instance, insists that distributional models are models of word meaning "as they are in the text" and not "in the head", so that DSMs should be considered primarily as computational models of meaning rather than "psychologically realistic model[s] of human semantic processing" (Sahlgren, 2008, pp. 134–135). Meanwhile, Günther et al. (2019) consider that DSMs stand in the long tradition of learning theories which postulate that humans are excellent at capturing statistical regularities in their environments. Yet, even if we are to agree with Günther et al. (2019), we must acknowledge that Sahlgren (2008) raises an important question: can distributional information found in corpora be considered representative of the type of distributional information grounding humans' conceptual representations in the first place?

[5] See also (Osgood, 1952, p. 228) very similar in spirit.

TABLE 2 | Corpora used to generate DSMs

Corpus	Word count	Details
OANC	17M	Open american national Corpus.[a] includes both spoken and written language, ranging from telephone and face-to-face conversations to letters, fiction, technical reports, newspapers or travel guides
WIKI07	19M	0.7% of the English wikipedia (WIKI) sampled across the entire dump
ACL	58M	Association for computational linguistics (ACL) anthology References corpus (Bird et al., 2008). Contains research papers in computational linguistics exclusively
WIKI2	53M	2% of the English wikipedia (WIKI) sampled across the entire dump. WIKI2 contains 12.5% of WIKI07
BNC	113M	British national Corpus.[b] includes both spoken and written language, ranging from informal conversations and radio shows to newspapers, academic books, letters or fiction
WIKI4	106M	4% of the English wikipedia (WIKI) sampled across the entire dump. WIKI4 contains 15% of WIKI07 and 100% of WIKI2
WIKI	2 600M	Full English wikipedia dump of January 20, 2019, generated and preprocessed (tokenize and lowercased) with WiToKit[c] based on wikiextractor[d] and polyglot (Al-Rfou et al., 2013). WIKI contains 100% of WIKI07, WIKI2 and WIKI4

[a] *https://www.anc.org/OANC/index.html*
[b] *http://www.natcorp.ox.ac.uk/*
[c] *https://github.com/akb89/witokit*
[d] *https://github.com/attardi/wikiextractor*

3.3.1 DSMs Are Not Grounded in Sensorimotor Experience

The first challenge faced by DSMs in their lack of *grounding* in sensorimotor experience of the real world, which makes them theoretically problematic as a sole account of meaning (e.g., De Vega et al., 2008; Wingfield and Connell, 2019). And indeed, Landauer and Dumais (1997) originally acknowledged that “to be more than an abstract system like mathematics, words must touch reality at least occasionally” (see p. 227). The problem is probably best illustrated by Harnad (1990) and his *Chinese/Chinese dictionary-go-round* example, itself an extension Searle’s *Chinese Room argument* (Searle, 1980): if one only had access to a Chinese/Chinese dictionary in order to learn the Chinese language, one would soon find themselves locked into a symbol/symbol merry-go-round that would render the task impossible (Harnad, 1990, pp. 339–340). As Glenberg and Mehta (2008) further note, no amount of statistical information can actually solve the problem of the circularity of definitions, if one cannot resort to alternative grounded modalities to understand what words actually *mean* (see p. 246).

By and large, such considerations raise the question of whether the type of *linguistic* distributional information found in text can be reasonably assumed to adequately mirror more *general* distributional information found in the world. As Connell (2019) puts it:

> Linguistic distributional statistics and simulated distributional statistics contain similar patterns, but do not directly reflect one another. In contrast to linguistic information, which comprises statistical regularities between word forms, simulated information encodes statistical regularities at the level of meaning due to the inclusion of situational context in simulated representations. A car, for instance, typically has wheels and a driver, operates on the road or street, and sometimes needs a service or repair. Objects, events, and other situational entities tend to occur together in the real world in ways that, through cumulative interactive experience, can give rise to statistical patterns of how referent concepts are distributed in relation to one another.

This question then extends to the question of the representativeness of *linguistic* distributional information in and of itself, and to whether what is found in standard DSM training corpora can be considered—both *quantitatively* and *qualitatively*—to constitute a representative sample of the type of linguistic distributional information humans are exposed to (Wingfield and Connell, 2019, pp. 8–11).

Yet, despite Connell’s concerns, several investigations have actually considered language to mirror the real world in ways that distributional information found in text could be assumed to reflect, in part or in full, distributional information grounded in sensorimotor experience (see, e.g., Barsalou et al., 2008; Louwerse, 2011). Be that as it may, what is important for our purpose here is not that distributional patterns found in corpora constitute *comprehensive* samples of distributional information grounding humans’ conceptual representations, but only that they condition the structural properties of the conceptual space in a plausible fashion (see more details in **Section 3.3.2**). Furthermore, insofar as the distributional hypothesis remains a hypothesis about *cascading variations—more* similar background experiences should entail *more* similar conceptual spaces—emphasis should be put on modeling plausible *differences* across distributional patterns speakers may be exposed to. We will return to that question in greater length in **Section 6**.

3.3.2 Can DSMs Nonetheless Model Conceptual Knowledge?

DSMs have historically been considered to model *conceptual* aspects of meaning, given how successful they prove to be at performing conceptual tasks such as *lexical similarity*, *priming* or *analogy* (see Westera and Boleda, 2019, §3.2). But can the vector for “cat” in a standard DSM really be considered to model the concept CAT when indeed it is only an abstraction over occurrences of the *word* cat and not over occurrences of *actual* cats? For Westera and Boleda (2019) it should not, and DSMs can at best be claimed to model *concepts of*

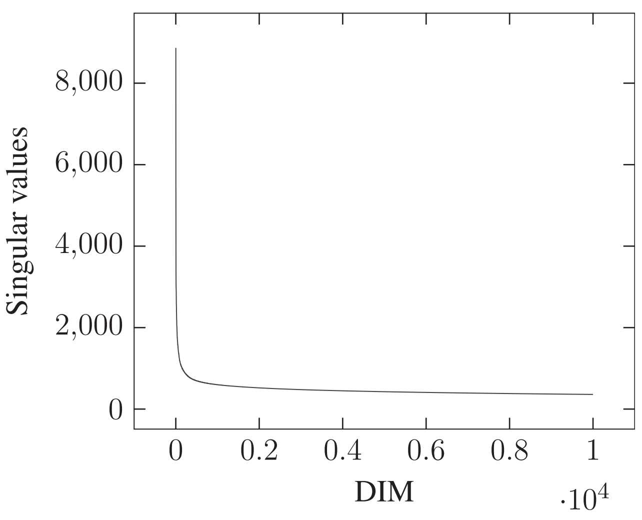


FIGURE 1 | Distribution of singular values across [0, 10,000] top dimensions for a PPMI-weighted count-based DSM generated on the full English Wikipedia (WIKI) corpus detailed in **Table 2**, with a window of 2.

words but definitely *not* concepts. And this distinction has its importance as, for them, one cannot expect relations that hold between concepts to necessarily hold between concepts of words. For example, the entailment relationship that may exist between CAT and ANIMAL may not necessarily hold between THEWORDCAT and THEWORDANIMAL.

Insofar as those considerations derive from the lack of grounding of DSMs previously detailed in **Section 3.3.1**, we will argue along the same lines. That is, we will not argue that DSMs provide comprehensive models of the conceptual space as a whole, but only that they provide satisfactory approximations for the purpose at hand. Our emphasis throughout this work being on the *structure* of the conceptual space—especially with respect to alignment—rather than, say, its cardinality, we remain mainly interested in the distribution of information across the dimensions of the DSM, and how that might be able to capture and reflect some structural properties of the conceptual space. In response to Westera and Boleda (2019), we will therefore say that, after all, concepts of words *are* concepts, so that even though DSMs were only able to model concepts of words, they could still be characterized as *subspaces* of a larger conceptual space, governed by similar constraints and structural properties: what matters here is not necessarily that, e.g., similar entailment relationships that hold between concepts also hold between concepts of words, but that *a* notion of entailment could be characterized in both the space of concepts and the subspace of concepts of words.

3.3.3 When DSMs Parallel Human Cognition

As we have previously mentioned, DSMs stand in the long tradition of learning theories which postulate that humans are excellent at capturing statistical regularities in their environments (Günther et al., 2019, p. 6). And in fact, as Connell and Lynott (2014) note, “natural languages are full of statistical regularities: words and phrases tend to occur repeatedly in similar contexts, just as their referents tend to occur repeatedly in similar situations” (see p. 395). Humans, as it appears, are sensitive to those regularities (e.g., Aslin et al., 1998; Solomon and Barsalou, 2004; Louwerse and Connell, 2011) which allows them to build conceptual representations from distributional knowledge (e.g., McDonald and Ramscar, 2001). Children, for example, are known to exploit statistical regularities in their linguistic environments, either via simple conditional probabilities when segmenting speech streams into words (Saffran et al., 1996), or via distributional patterns when acquiring syntactic knowledge (Redington et al., 1998).[6]

Jenkins (1954) originally proposed a summary of the whole lexical acquisition process: “intraverbal connections arise in the same manner in which any skill sequence arises, through repetition, contiguity, differential reinforcement” (see p. 112). Since then, several research have argued that the learning of associations between stimuli is driven by *contingency* rather than *contiguity* (Rescorla and Wagner, 1972).[7] As Rescorla (1968) details, the notion of contingency differs from contiguity in that it takes into account not only what *is* there but also what *is not* in the form of conditional probabilities. In essence, the notion of contingency characterizes the *informativity* of a given stimuli. For Günther et al. (2019), PPMI-based DSMs directly follow such learning theories as they indeed encode *mutual information* between words and contexts, that is, their respective informativity, rather than raw word-context co-occurrence count.

A crucial aspect of DSMs is that they follow the emergentist approach to cognitive development (e.g., Elman et al., 1996) and conceptual representations (e.g., Rogers and McClelland, 2004) in considering that long-term knowledge is an emergent representation abstracted across multiple experiences. Within the emergentist family of connectionist models, there is no real distinction between knowledge of something and knowledge of the contexts in which that thing occurs, and several implementations have historically been proposed to show how a conceptual representation could be abstracted from contextual experience (e.g., Elman, 1990; Elman, 1993; Altmann, 1997).

For Jones et al. (2015) both the connectionist and the distributional approaches have in common to hypothesize the existence of a *data reduction* mechanism that enables focusing on important statistical factors that are constant across contexts while throwing away factors that are idiosyncratic to specific contexts (see p. 240). Landauer and Dumais (1997) argued early on that the dimensionality reduction step in the DSM generation pipeline could model the transition from episodic to semantic memory,[8] formalized as the generalization of observed concrete instances of word-context co-occurrences to higher-order representations potentially capturing more fundamental and conceptual relations (see p. 217). The idea that DSMs could provide computational models of semantic memory can also be found in (McRae and Jones, 2013; Jones et al., 2015).

[6] See also (Saffran, 2003; Smith and Yu, 2008; Aslin and Newport, 2012; Hall et al., 2018).

[7] Although see maybe (Papini and Bitterman, 1990) for a counter-argument.

[8] Episodic memory is assumed to contain memory of autobiographical events while semantic memory is assumed to be dedicated to generalized memory not linked to specific events (Tulving, 1972).

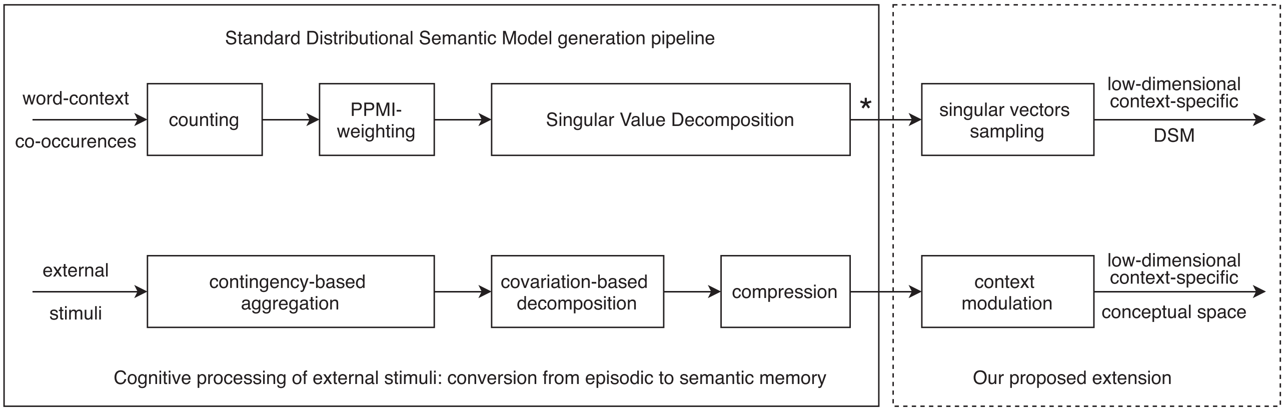

Standard Distributional Semantic Model generation pipeline
word-context
co-occurences
counting
PPMI-weighting
Singular Value Decomposition
*
singular vectors sampling
low-dimensional context-specific
DSM
external
stimuli
contingency-based aggregation
covariation-based decomposition
compression
context modulation
low-dimensional context-specific
conceptual space
Cognitive processing of external stimuli: conversion from episodic to semantic memory
Our proposed extension

FIGURE 2 | Parallel between the standard DSM generation pipeline (upper part) and human cognition (lower part). The left part of the diagram details the cognitive processing of external stimuli converting episodic memory to semantic memory, while the right part is our proposed extension for modeling ad-hoc coordination (detailed in **Section 4.1**). Note that the standard output of a traditional DSM generation pipeline (see *) is a low-dimensional matrix of dimension $k \approx 300$ made of top variance-preserving components. As we detail in **Section 4.1**, we replace this by top $k = 10\,000$ components from which we sample a subset of singular vectors.

Another important assumption made with respect to this compression mechanism is that it relies on a form of covariation-based decomposition of the previously aggregated stimuli. As such, it operates in a similar fashion than Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) in being able to structure and organize latent information based on variance: broad, high-order distinctions come first before more fine-grained ones (Rogers and McClelland, 2004; Jones et al., 2015). This assumption is supported by empirical evidence showing that children acquire concepts through progressive differentiations: 18-months-olds first develop global conceptual categories such as *animals*, *vehicles*, *plants*, *furniture* and *kitchen utensils* before being able to operate high-constrat basic-level distinctions among those categories by 30 months, and ultimately learning to operate low and moderate basic-level contrasts among those categories later one (Mandler et al., 1991).

Now Glenberg and Mehta (2008) argued that covariation among words is not sufficient to characterize meaning, and showed that participants failed to rely on covariance structure to, e.g., classify unnamed features for familiar domains. Yet, this does not mean that covariation cannot be used as a proxy to capture certain conceptual properties such as lexical similarity. Again, the fact that concepts cannot be characterized by covariation alone does not make it useless. Note again here, as Landauer and Dumais (1997) have stressed before, that we do not need to consider SVD to constitute *the* cognitive mechanism used by humans to perform data compression. We can just assume that the brain uses some sort of dimensionality reduction mechanism *akin* to SVD in order to create abstract conceptual representations by favoring high covariance structure while eliminating idiosyncrasies.

In short, the standard DSM generation pipeline can be considered to parallel human cognition via three specific processes (see **Figure 2**): 1) *contingency-based aggregation* of distributional information through word-context co-occurence counting and PPMI-weighting; 2) covariation-based *decomposition* through Singular Value Decomposition; and 3) *compression* through dimensionality reduction of the SVD matrix.

4 MODEL AND EXPERIMENTAL SETUP

4.1 Modeling Coordination as Singular Vectors Sampling

Recall the "dog" example of Connell and Lynott (2014) previously introduced in **Section 2.3**: imagine yourself discussing *dogs* with someone who has only ever encountered dogs the size of a German shepherd while you have only encountered dogs the size of a chihuahua. At the beginning of the conversation, those differences across background experiences could translate as differences across your respective similarity judgments: assuming here for the sake of the argument that all similarity judgments are solely based on a *size* feature, you may think that DOG is more similar to CAT or even to MOUSE than to BEAR, while your interlocutor may think the opposite. Yet, provided that you talk long enough, you and your interlocutor may somehow accommodate those discrepancies across your respective background experiences and update your conceptual representations of *dogs* accordingly. This may in turn translate as cascading updates in your similarity judgments, and at the end of the conversation you may then both consider DOG to be more similar to CAT than to BEAR, and to be more similar to BEAR than to MOUSE.[9]

In this work we propose to characterize superficial alignment during ad-hoc coordination as the cooperative act of *aligning*

[9] We are not arguing here that similarity judgments are always necessarily "feature-based" or that there could exist more prominent features systematically influencing similarity judgments. We only provide this example for illustrative purposes in order to give the reader a better intuition of our sampling algorithm's underlying logic. Nonetheless, our examples remains grounded in empirical evidence which show, e.g., that novices tend to judge similarity based on superficial or surface features, whereas experts rely on deeper underlying principles (Chi et al., 1981). So in fact our example is not completely implausible as differences in knowledge grounded in differences across background experiences could perfectly translate as differences in similarity judgments: a biologist could be more inclined to consider that Cat is more similar to Tiger than to Dog on the ground of their being both part of the *felidae* family, while someone else, say a child, unaware of this sort of classification, could consider Dog and Cat to be more similar on the basis of their being both of similar *shape* or *size*.

lexical similarity judgments on a limited set of word pairs. In practice, we propose to model coordination with DSMs as *singular vectors sampling*: we modify the standard DSM generation pipeline by replacing the ill-motivated variance-preservation bias described in **Section 3.2** by an explicit coordination bias, sampling the set of d singular vectors which maximize the correlation with a particular lexical similarity dataset. The core assumption underlying our sampling algorithm is that it is actually possible for DSMs to capture different kinds of semantic relations from the same corpus, so that rather than generating a *single* meaning space from the PPMI matrix, a *collection* of possible meaning spaces could coexist within the same set of data. The collocates of *cat*, for instance, could provide enough information to characterize it as similar to *tiger* on the one hand (i.e., having a neighborhood of ontologically-related words), or to *meow* on the other hand (i.e., having a neighborhood of generically related words), and be aggregated in different dimensions during the factorization step. This assumption will be supported later on by our experimental results showing that DSMs relying on our sampling algorithm rather than the variance-preservation bias can indeed perform significantly better on several lexical similarity datasets, as different dimensions encode different semantic phenomena (see **Section 5.1**).

In practice, given that the rank r of the sparse PPMI matrix is usually well beyond a manageable order of magnitude ($r >$ 100,000) to explore all possible subsets in U, we propose a sampling algorithm to efficiently sample only a limited number of subsets of singular vectors in U. Our sequential (seq) sampling algorithm works in two passes:

(1) add: during the first pass, the algorithm iterates over all singular vectors and selects only those that increase performance on a given dataset;
(2) reduce: during the second pass, the algorithm iterates over the set of added singular vectors and removes all those that do not negatively alter performance on the given dataset.

The structure of the algorithm, especially the presence of the reduce step, is motivated by the presence of many complex semantic redundancies across singular vectors from the point of view of fitting a particular meaning space, so that adding a particular singular vector to a set pre-existing ones may make some of them redundant.

Additionally, and for computational efficiency, we reduce the number of singular vectors under consideration by sampling over the top-k singular vectors only, with k = 10,000.[10] The algorithm can be run through multiple iterations, and may iterate over singular vectors in linear or shuffled order (of singular value). We apply 5-fold validation and report scores averaged across test folds, with the corresponding standard error. We define performance on a given similarity dataset as both the Spearman correlation *and* the Root Mean Square Error (see **Eq. 6**) computed on a set of word pair similarities. That is, the sampled models have to align *both* the ranking *and* the absolute similarity values of the set of word pairs with that of the dataset. This feature modeling choice is motivated by preliminary results on k-fold validation showing a tendency to overfit when performance metric is restricted solely to Spearman correlation.[11]

In effect, our model approximates coordination as context modulation, where context modulation is understood as the act of accommodating past experienced contexts to the specific context of the discussion. Indeed, several research have shown that dimensions in DSMs capture different contexts in which words are used (e.g., Griffiths et al., 2007, p. 221) so that, in fact, the process of singular vectors sampling is tantamount to context selection and aggregation. The main benefit of our approach is that it allows us to model cascading conceptual modulation across the *entire* conceptual space. Since latent singular vectors condition the content of *all* semantic representations, sampling a set of singular vectors will not just impact the representations of the lexical items being aligned, but actually the entire conceptual space. Moreover, this mechanism of singular vectors sampling is theoretically very convenient as it relieves us from having to formulate explicit assumptions regarding the latent structure of the conceptual space: cascading modulation will always be conditioned on latent interdependencies which are grounded in shared contextual aggregates across semantic representations.

Note, however, that we do not model conceptual update, neither *during* nor *after* coordination. As a matter of fact, since we assimilate coordination to the act of *accommodating existing knowledge* to the situation at hand, we do not actually need to update the original PPMI matrix, which relieves us from having to formulate a theory about how conceptual update could and should proceed in such situations. Since our main purpose throughout this study is to investigate the dynamics of alignment during ad-hoc coordination, we can actually focus on an approximation of the coordination process between any two arbitrary points in time. Similarly, we do not model online coordination at every step of the process—such as conceptual update occurring at every utterance during real-time communication—as we do not need this level of granularity for the purpose at hand. Once again, this should be seen as an opinionated modeling decision rather than a limitation of our model.[12]

Finally, we exclusively focus here on count-based DSMs given that, as we have seen in **Section 3.3** their generation pipeline nicely

[10] Our choice of k = 10 000 is questionable given that we previously showed in **Table 1** that it could at best retain 72% of the total energy. It is primarily motivated by computational considerations and the necessity to maintain acceptable overall computing time. As we show in **Section 5.1** it appears to be a reasonable compromise given than 90% of the sampled dimensions on all our DSMs and across all our lexical similarity datasets remain below the 8,000th dimension.

[11] Note that in order to minimize interferences with reported results, we perform pre-validation on the MEN and SimLex datasets only, using DSMs generated exclusively from the WIKI corpus. Full details regarding this point are made available in the **Supplementary Material**.

[12] On practical matters, note that a rich literature exists on incremental SVD update (e.g., Businger, 1970; Bunch and Nielsen, 1978) so that our particular implementation would not necessarily constitute an obstacle to modeling online coordination: see (Gentle, 2009) for a comprehensive introduction to the topic. Brand (2003, 2006) has notably proposed an algorithm for incrementally adding, removing and updating rows and columns in the SVD matrix that could prove particularly useful for that purpose.

parallels the functioning of human cognition. Moreover, they provide more transparent, explainable and modular models in comparison to their prediction-based counterparts, which makes it easier to operate cognitively-motivated modeling modifications. It appears difficult indeed to transpose our proposed approach to prediction-based DSMs as-is. The singular vectors sampling mechanism could probably be replaced by a kind of post-processing technique akin to what Mu and Viswanath (2018) have used for instance as a way to somehow bypass the variance-preservation bias. But those postprocessing techniques have yet to be formalized for our purpose and one would loose the benefits of sampling on dimensions that explicitly capture context aggregates. Not to mention additionally that those postprocessing techniques usually rely on linear transformations that are sort of “one-shot” and cannot necessarily easily be made to function incrementally.

4.2 Measuring Conceptual Alignment via Matrix Transformation

In the previous section we proposed to characterize superficial alignment during ad-hoc coordination as the cooperative act of aligning lexical similarity judgments on a limited set of word pairs. Recall from **Section 2.4**, however, that we stressed the need to distinguish superficial from structural alignment when investigating the socialization hypothesis, as the two do not necessarily go hand-in-hand. We argued more specifically in favor of a notion of conceptual similarity that could quantify the overall structural similarity between any two conceptual spaces, while potentially tolerating high degrees of local and superficial variability.

In this section we therefore propose to model structural similarity between two DSMs as the minimized Root Mean Square Error (RMSE; **Eq. 6**) between them. DSMs are first aligned using *absolute orientation with scaling* (see Algorithm 1 below from Dev et al., 2018, originally Algorithm 2.4 in their paper) where the optimal alignment is obtained by minimizing the sum of squared errors under the Euclidian distance between all pairs of common data points, using linear transformations—rotation and scaling—which do not alter inner cosine similarity metrics and hence preserve measures of pairwise lexical similarity.

Algorithm 1: | Absolute orientation with scaling AOS(A, B)

```
Compute the sum of outer products $H = \sum_{i=1}^{n} b_i^T a_i$
Decompose $[U, S, V^T] = \mathrm{svd}(H)$
Build rotation $R = UV^T$
Rotate $\tilde{B} = BR$ so each $\tilde{b}_i = b_i R$
Compute scaling $s = \sum_{i=1}^{n} \langle a_i, \tilde{b}_i \rangle / \|\tilde{B}\|_F^2$
return $\breve{B}$ as $\breve{B} \leftarrow s\tilde{B}$ so for each $\breve{b}_i = s\tilde{b}_i$
```

The Root Mean Square Error (RMSE) between the two matrices A and $\breve{B}$ is then given by:

$$RMSE(A, \breve{B}) = \sqrt{\frac{1}{|A|} \sum_{i=1}^{|A|} \|a_i - \breve{b}_i\|^2} \qquad (6)$$

Note that due to floating point approximations, our computed RMSEs are not symmetric, so that $RMSE(A, \breve{B}) \neq RMSE(\breve{A}, B)$, with $\breve{B} = AOS(A, B)$ and $\breve{A} = AOS(B, A)$. To alleviate this problem, we always report the averaged RMSE: $\overline{\mathrm{RMSE}} = 1/2\,[\mathrm{RMSE}(A, \breve{B}) + \mathrm{RMSE}(\breve{A}, B)]$.

Our notion of structural similarity follows alignment-based models (Goldstone and Son, 2012, p. 165) in that it attempts to place elements of the two DSM matrices in correspondence with one-another via a set of structure-preserving operations, and therefore does not measure a raw comparison between them. The underlying methodology has been widely used in computational linguistics to align DSMs across languages (e.g., Mikolov et al., 2013b) although it is to be distinguished from other alignment-based approaches in the field which apply potentially non-cosine-preserving linear transformations (e.g., Tan et al., 2015). Such methodologies can also be found in neuroscience with the *hyperalignment* approach put forth by Haxby et al. (2011) which proposes to align patterns of neural response across subjects using linear transformations—namely rotations and reflections—minimizing the Euclidian distance between two sets of paired vectors, in order to abstract away the intrinsic variability of voxel spaces across subjects. The underlying logic is always the same: two models can be transformationally equivalent although they may not appear similar in absolute. Aligning the coordinate system or the basis of two vector spaces, for instance, can uncover measures of relative similarity between two models that otherwise appear radically different when comparing only their original respective coordinate values.

Recall also from **Section 4.1** that we proposed to model superficial alignment during coordination with DSMs as singular vectors sampling, with the benefits thereby of being able to model cascading conceptual modulation across the entire conceptual space. The question that arises, then, is if, as defined, superficial alignment will necessarily entail structural alignment. That is, will maximizing the Spearman correlation on a lexical similarity dataset using our singular vectors sampling algorithm on two DSMs generated from two distinct corpora in turn lower the RMSE between them. We report our results on the matter in **Section 5.2**.

It is important to note here, however, that the connection between our characterizations of superficial and structural alignment are not necessarily obvious. Indeed, our notion of structural similarity satisfies the requirements detailed in **Section 2.4** in that it can indeed tolerate high degrees of local and superficial variability: since the RMSE-based structural similarity measures absolute distances between points in space, it is insensitive to relative measures of semantic proximity, unlike what is expected from correlations with lexical similarity datasets. Naturally, if two DSMs have a null RMSE, they will produce identical similarity judgments on a set of word pairs. But the slightest deviation from 0 can have unpredictable consequences depending on the configuration of the space. So in fact, our model makes it possible for any two DSMs to behave very differently with respect to lexical similarity while actually being well aligned structurally (and conversely) following thereby the position of Wachsmuth et al. (2013) detailed in **Section 2.4**.

4.3 Experimental Setup

We generate PPMI-weighted DSMs using a window of size 2 from seven different corpora detailed in **Table 2**. All corpora are lowercased and tokenized with Polyglot (Al-Rfou et al., 2013). All Wikipedia subsets are generated by sampling the WIKI corpus at the sentence level. Corpora are chosen so as to provide pairs of comparable size (OANC and WIKI07; ACL and WIKI2; BNC and WIKI4) covering different domains and/or different genres (see details in **Table 2**). Note that our point here, as we have previously detailed in **Section 3.3.1**, is not to model plausible individual speakers, but plausible *differences* across background experiences. What is important therefore is not that corpora be produced by individual speakers, or even characterize the linguistic experience of individual speakers, but that the differences across their linguistic distributional patterns model plausible differences of background experiences. We return to this question in more details in **Section 6**. We therefore select corpora which we assume to characterize quite different linguistic distributional patterns: ACL for instance covers exclusively research papers in computational linguistics, while OANC and BNC both include spoken and written language from different genres (newspapers, fiction, technical reports, travel guides, etc.).

For word similarity datasets, we rely on MEN (Bruni et al., 2014), SimLex-999 (hereafter SimLex (Hill et al., 2015); and SimVerb-3500 (hereafter SimVerb; Gerz et al., 2016). MEN is a relatedness dataset containing a list of 3,000 word pairs with a strong bias toward concrete concepts; while SimLex intends to encode *similarity* rather than *relatedness* for 999 word pairs, and provides a more balanced account between *concrete* and *abstract* concepts. Words that have high relatedness in MEN may have low similarity in SimLex. For example, the pair "chicken-rice" has a similarity score of 0.68 in MEN and 0.14 in SimLex. Following previous claims and standard linguistic intuitions, the relatedness dataset MEN should be only weakly compatible with the similarity dataset SimLex: one expresses topical association (i.e. *cat* and *meow* are deemed related) while the other expresses categorical similarity (i.e. *cat* and *dog* might be considered similar in virtue of being members of the same category). Thus, those datasets encode possibly incompatible semantic constraints and it is theoretically impossible to perfectly fit both the meaning spaces they encode with a single DSM. Those two datasets therefore allow exploring our approach across two distinct coordination situations. The third dataset, SimVerb, is a similarity dataset consistent with SimLex, but focusing on verb meaning and providing 3,500 word pairs. Although theoretically compatible with the notion of similarity encoded in SimLex, it focuses on different semantic categories and as such on a potentially different domain with distinct semantic constraints. Given that we rely on MEN and SimLex for pre-validation of our sampling algorithm (see **Section 4.1**) we add SimVerb as an additional dataset to further check the robustness of our results.

Mincount hyperparameters are set so as to maximize lexical coverage on all similarity datasets while maintaining reasonable overall computing time. We choose a mincount of 2 for OANC, WIKI07, ACL, WIKI2, BNC and WIKI4 and 30 for WIKI. Lexicons are aligned across all DSMs *after* the SVD computation and we obtain a MEN coverage of 93.0% (2,817 pairs out of 3,000), a SimLex coverage of 99.5% (994 pairs out of 999) and a SimVerb coverage of 94.91% (3,322 pairs out of 3,500).

We compute p values on each test fold using a Steiger's test (Steiger, 1980)[13] following (Rastogi et al., 2015). We consider as the null hypothesis the fact that two models perform identically on a given lexical similarity dataset. We then combine all p values for a given k-fold using the weighted harmonic mean (see Wilson, 2019) treating folds as dependent tests, and report a single p value per k-fold.

Finally, we make our code available for replication at https://gitlab.com/akb89/avoiding-conflict.

5 RESULTS

5.1 No Variance-Preservation Bias Means Better Superficial Alignment

We first report the performance of our seq sampling algorithm described in **Section 4.1** against PPMI-weighted count-based (TOP) models reduced by selecting the top n singular vectors in the SVD matrix, with ($\alpha = 1$) or without ($\alpha = 0$) singular values. In order to provide a completely fair comparison across models, we generate for each fold a specific TOP model with the exact same number of dimensions n than the one sampled by our seq algorithm for that particular fold. We similarly compute the statistical significance of the difference of performance between the SEQ and the TOP models *per fold*. We then report a single Spearman correlation per model, corresponding to the *mean* and *standard error* across all 5-folds, and report a single statistical significance score, computed as the harmonic mean of the p values across five folds, as previously detailed in **Section 4.3**.

Our results show that replacing the traditional variance-preservation bias with our sampling algorithm leads to near-systematic improvements on all corpora and across all similarity datasets (see **Table 3**). The detrimental effect of variance-preservation is first exemplified when comparing DSMs with singular values ($\alpha = 1$) to those without ($\alpha = 0$), an effect originally noted by Caron (2001) and also discussed by Levy et al. (2015). This detrimental effect is then further exemplified by introducing our sampling algorithm and proves most salient on the ACL corpus, with a 17 points increase in performance on MEN, a 13 points increase on SimLex, and a 12 points increase on SimVerb, all statistically significant ($p < 0.01$).

Explicitly sampling singular vectors leads to an even more interesting observation: *different dimensions encode different semantic phenomena*. Contrary to what was originally argued in (Schütze, 1992, p. 794), all singular vectors are not necessarily meaningful to discriminate particular patterns of word similarities. For example, the semantic phenomenon of *relatedness* encoded in MEN is characterized by a different sampling pattern than the *similarity* phenomenon encoded in either SimLex or SimVerb (see **Table 4**). Overall, MEN is characterized by *higher* singular vectors, when SimLex and

[13] As implemented by Philipp Stinger: https://github.com/psinger/CorrelationStats/blob/master/corrstats.py

TABLE 3 | Spearman correlations on MEN, SimLex and SimVerb for DSMs generated from different corpora.

Model	α	WIKI07	OANC	WIKI2	ACL	WIKI4	BNC
				MEN			
TOP	1	0.48 ± 0.01	0.50 ± 0.01	0.53 ± 0.02	0.25 ± 0.03	0.54 ± 0.01	0.61 ± 0.01
TOP	0	0.56 ± 0.01	0.59 ± 0.01	0.61 ± 0.01	0.34 ± 0.02	0.62 ± 0.01	0.69 ± 0.01
SEQ	-	**0.60** ± 0.01	**0.64** ± 0.01	**0.66** ± 0.01	**0.51** ± 0.01	**0.69** ± 0.02	**0.74** ± 0.00
p value		0.0023	0.0003	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
ndim		186 ± 5	195 ± 6	200 ± 3	300 ± 9	215 ± 6	161 ± 6
				SimLex			
TOP	1	0.20 ± 0.04	0.18 ± 0.02	0.24 ± 0.02	0.11 ± 0.06	0.25 ± 0.02	0.27 ± 0.03
TOP	0	0.24 ± 0.03	0.22 ± 0.02	0.26 ± 0.02	0.14 ± 0.05	0.27 ± 0.02	0.32 ± 0.03
SEQ	-	0.25 ± 0.03	0.19 ± 0.04	0.30 ± 0.03	**0.27** ± 0.03	**0.38** ± 0.02	**0.41** ± 0.02
p value		0.3802	0.0906	0.0646	0.0001	0.0010	0.0056
ndim		184 ± 12	240 ± 9	196 ± 12	221 ± 5	224 ± 6	201 ± 10
				SimVerb			
TOP	1	0.08 ± 0.02	0.07 ± 0.02	0.11 ± 0.03	0.07 ± 0.01	0.12 ± 0.01	0.16 ± 0.01
TOP	0	0.13 ± 0.01	0.13 ± 0.02	0.15 ± 0.03	0.11 ± 0.01	0.17 ± 0.01	0.22 ± 0.02
SEQ	-	**0.20** ± 0.03	0.19 ± 0.02	**0.21** ± 0.03	**0.23** ± 0.01	**0.25** ± 0.01	**0.29** ± 0.01
p value		0.0019	0.0216	0.0001	0.0015	0.0043	0.0015
ndim		290 ± 17	185 ± 12	317 ± 18	267 ± 13	376 ± 11	331 ± 12

All models are PPMI-weighted count-based models generate with a window size of 2. SEQ models are reduced via our seq algorithm detailed in ***Section 4.1****, while TOP models are reduced by selecting the top n = ndim singular vectors from the SVD matrix, with ndim corresponding for each fold to the number of dimensions sampled by the SEQ model on that fold. All results are averaged across test folds applying 5-fold validation, after taking the best of 10 shuffled runs. Bold results indicate statistically significant differences ($p < 0.01$) between SEQ and TOP ($\alpha = 0$) models.*

TABLE 4 | Average mean, median and 90-th percentile of sampled dimensions indexes on MEN, SimLex and SimVerb for 10 shuffled runs in seq mode.

	MEN			SimLex			SimVerb		
	Median	Mean	90%	Median	Mean	90%	Median	Mean	90%
WIKI07	196 ± 12	576 ± 58	995 ± 237	612 ± 46	1,917 ± 107	6,314 ± 325	564 ± 45	1768 ± 98	6,227 ± 315
OANC	172 ± 9	567 ± 64	1,022 ± 168	677 ± 70	2,003 ± 92	6,499 ± 200	672 ± 68	2,210 ± 97	7,371 ± 200
WIKI2	220 ± 13	462 ± 48	917 ± 89	606 ± 35	1,218 ± 64	3,091 ± 242	586 ± 26	1,188 ± 60	2,847 ± 253
ACL	586 ± 15	1,233 ± 43	3,201 ± 178	935 ± 80	2,289 ± 106	7,376 ± 212	717 ± 47	1852 ± 79	6,012 ± 330
WIKI4	270 ± 11	532 ± 35	1,120 ± 59	662 ± 27	1,177 ± 50	2,635 ± 209	721 ± 37	1,297 ± 67	3,100 ± 260
BNC	163 ± 8	419 ± 48	651 ± 84	439 ± 22	969 ± 67	2,285 ± 291	518 ± 21	980 ± 41	2,254 ± 83

SimVerb are characterized by *lower* and more latent ones, which could explain the historical success of variance-based DSMs at capturing semantic relatedness rather than similarity. Moreover, our results show that models generated from different corpora will distribute information differently across their singular vectors, as shown per the variations of sampling patterns within identical similarity datasets displayed in **Table 4**: ACL-based DSMs for instance encode MEN much more latently in comparison to other corpora ($\overline{dim_i} = 1,233 \pm 43$) which explains the originally low performance on MEN of the variance-based DSM generated from ACL (see TOP scores for ACL in **Table 3**). In short, the information necessary to characterize a particular semantic phenomenon may actually be present (at least to some extent) in a given corpus, but not actually distributed in the top components of the SVD, calling once again into question the pertinence of the variance-preservation bias.

5.2 Better Superficial Alignment Does Not Mean Better Structural Alignment

Results of **Section 5.1** show that explicit singular vectors sampling on MEN, SimLex and SimVerb leads to increased superficial alignment across datasets, and that the sampled singular vectors do *not* systematically correspond to the top components of the SVD. Still, would those specific sampling patterns also improve structural alignment between DSMs by lowering their RMSE? *Probably not.* To prove our point, let us plot the evolution of RMSE across bins of 250[14] consecutive singular vectors, for corpora of same size but different domains (**Figure 3**) and different size but similar domains (**Figure 4**).

[14] Corresponding to the rough average number of singular vectors sampled across models and datasets in **Table 3**.

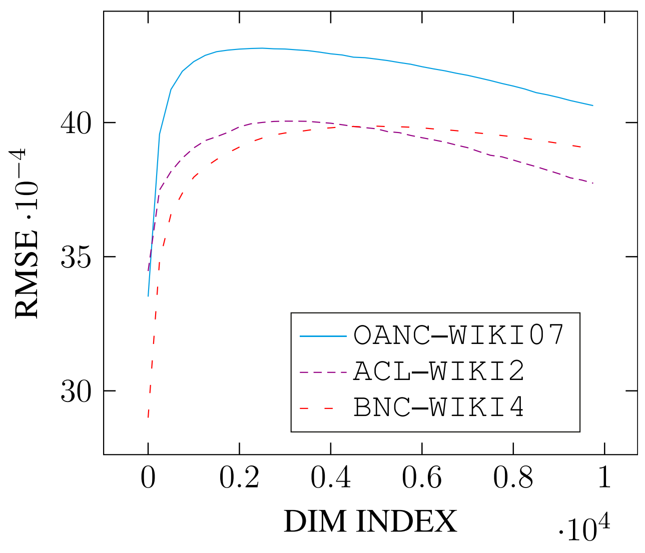


FIGURE 3 | Evolution of RMSE for aligned bins of 250 consecutive singular vectors sampled across [0, 10 000] for aligned corpora of different domains but similar size.

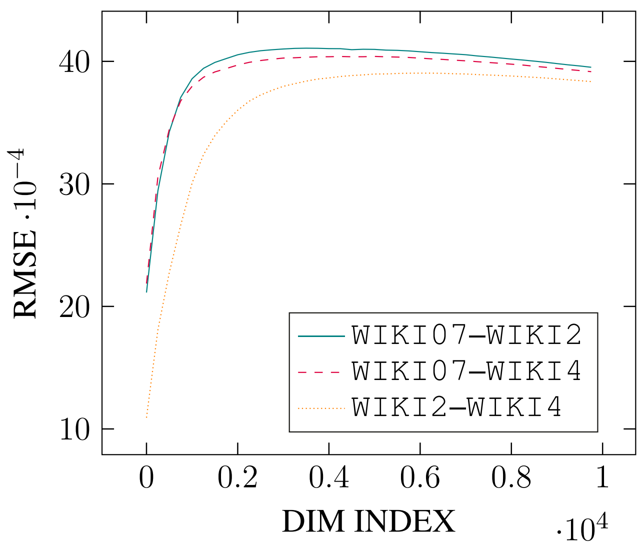


FIGURE 4 | Evolution of RMSE for aligned bins of 250 consecutive singular vectors sampled across [0, 10 000] for aligned corpora of similar domains but different size.

What those plots show first is the ability for our structural similarity metric to capture the intuition of *similar domains* across corpora: plots displaying the evolution of RMSE computed over pairs of models of partly overlapping Wikipedia samples follow much more similar trends than plots over pairs of models from different domains (compare gaps between plots across **Figure 4** and **Figure 3**). What they show next, however, in that the RMSE is minimal for the top 250 components of the SVD and that it rapidly increases then. Therefore, any sampled set of 250 non-top singular vectors such as those reported in **Table 4** will necessarily obtain a higher RMSE in comparison. In other words, increasing superficial alignment will necessarily decrease structural similarity.

5.3 Beyond Structural Alignment: Agreement vs. Compatibility

Figure 3 and **Figure 4** exhibit a similar global pattern across aligned models: to minimize the RMSE, singular vectors can be sampled via the very top or the much more latent part of the SVD. Those two parts of the SVD, however, capture quite different information: more systematic information about language for the top components, and more idiosyncratic information regarding the corpus at hand for the more latent components. This phenomenon can be quantified by plotting the absolute Pearson correlation between pairs of singular vectors sampled across two DSMs (see **Figure 5**): top components have a correlation value closer to 1 ~(log ≈ 0) although it rapidly decreases as we move toward more latent singular vectors.

And yet, as we plot the evolution of the RMSE as a function the Pearson correlation, averaged on bins of 30 consecutive singular vectors sampled across [0, 10 000], we do not observe a linear curve: that is, alignment does not get more and more difficult as the Pearson correlation decreases, but reaches a peak before significantly diminishing again (see **Figure 6**). This further illustrates a fundamental property of our alignment-based notion of similarity: two given models may be aligned if they both have *similar* components, but also if they have *dissimilar* components, provided that those components do not *conflict*. Notions of *agreement*, *compatibility* and *conflict* can be defined via the absolute Pearson correlation as described in **Figure 6**: maximal agreement is given by an absolute Pearson correlation of 1, and maximal compatibility is given by an absolute Pearson correlation of 0. In between, conflict increases as the absolute Pearson correlation goes down from full agreement to the peak of disagreement which maximizes the RMSE, then decreases again until it reaches maximal compatibility. Concretely, the peak of disagreement will correspond to sampling patterns that maximize structural dissimilarity between conceptual spaces, although this may not necessarily translate as superficial dissimilarity and explicit conflict between speakers during conversation, for reasons explained in **Section 2.4**. Note, moreover, that agreement and compatibility are defined on different domains: agreement is only defined rightward of the peak of disagreement, while compatibility is only defined leftward of the peak. Therefore, two speakers in full agreement cannot be said to have *incompatible* conceptual spaces.

A concrete example detailing the underlying mathematics of agreement and compatibility is given in **Eq. 7**: both matrix B and C can be aligned with matrix A when using our alignment algorithm, with a near-null RMSE ($< 10^{-15}$). Yet, both matrices have quite different Pearson

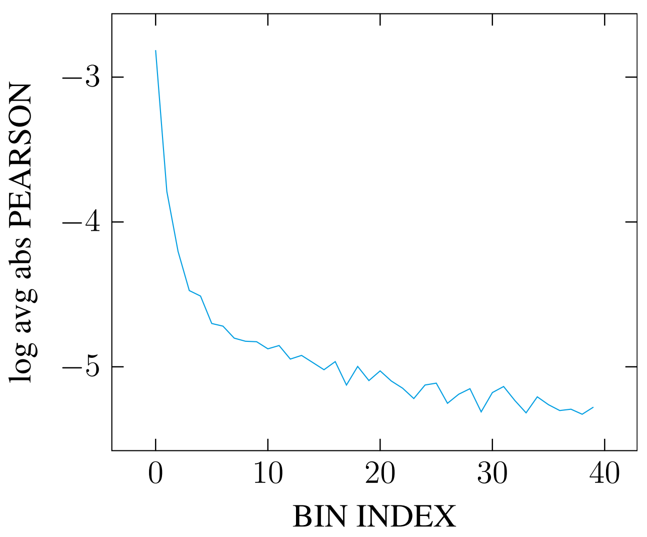


FIGURE 5 | Evolution of the log of the average absolute pariwise Pearson correlation between singular vectors for bins of 250 sampled across [0, 10 000] on OANC and WIKI07.

correlations: B's and A's elements have similar values and therefore A and B's column vectors have a pairwise Pearson correlation of 1, while A and C's pairwise Pearson correlation is merely at 0.3.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} .9 & 0 & 0 & 0 \\ 0 & .9 & 0 & 0 \\ 0 & 0 & .9 & 0 \\ 0 & 0 & 0 & .9 \end{bmatrix} \qquad (7)$$
$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

This phenomenon directly relates to the "dog" example of Connell and Lynott (2014) previously detailed in **Section 2.3**, which showed how alignment may not always equate agreement but sometimes mere compatibility between conceptual representations: speakers holding marginally identical conceptual representations—in this case widely differing representations of prototypical dogs *size*—can still be assumed to understand one-another, especially if disagreement pertains to aspects of conceptual knowledge that are irrelevant to the conversation at hand. Our experimental results support the idea that such considerations also extend to conceptual *spaces* and notions of structural similarity: widely differing aggregates of contextual experience captured by singular vectors can still sometimes provide a solid basis for structural alignment. Our characterizations of notions of structural agreement and compatibility, however, are more flexible than previous ones, in that they notably do not require a form of explicit, lexicalized, "feature-based" interpretation of what they entail. In our case, they can be defined in a more systematic fashion as a form of latent structural property of the conceptual space with respect to alignment.

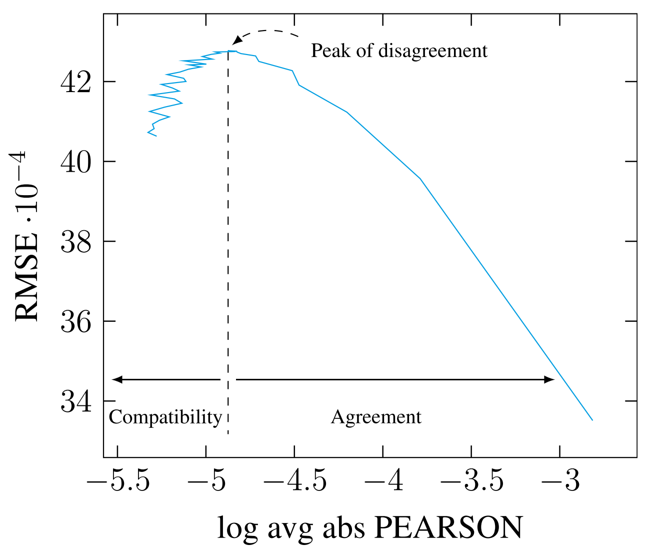


FIGURE 6 | Evolution of RMSE with log of average absolute Pearson correlation for aligned bins of 250 consecutive singular vectors sampled across [0, 10 000] on OANC and WIKI07.

6. DISCUSSION

6.1. Why Is Compatibility Relevant Anyway?

Why should we care about compatibility in the first place? After all, **Figures 3**, **4**, and **6** combined show that the RMSE is significantly lower in the agreement zone than in the compatibility zone, especially for the top components of the SVD. Why should speakers striving to align their conceptual spaces, then, not end up sampling those top components, and only those top components? The answer to that question will depend on how many singular vectors we can reasonably assume to be sampled during a realistic coordination setting. Because the RMSE is certainly lowest for the top components of the SVD, but those top components are actually not that many: after the first 250 singular vectors, the RMSE then significantly increases across all corpora in a systematic fashion.

And indeed when looking at it more closely, the compatibility zone appears to include *many more* singular vectors than the agreement zone. Our results show indeed that the peak of disagreement is located roughly at $d = 2,175$ for OANC–WIKI07, $d = 2,850$ for ACL–WIKI2, and $d = 4,750$ for BNC–WIKI4, out of 10,000 singular vectors in total. Yet the comparison does not stop there as the location of the peak of disagreement alone does not guarantee that singular vectors sampled from the agreement zone will systematically lead to lower RMSE compared to singular vectors sampled from the compatibility zone. As a matter of fact, numbers drop even further then: only about 225 singular vectors of the 2,175 that are in the agreement zone of OANC–WIKI07 can lead to a lower RMSE than the lowest RMSE of the compatibility zone. For ACL–WIKI2, the corresponding number is about 250 out of 2,850, and for BNC–WIKI4, 1,400 out of 4,750.[15]

[15] All those numbers were computed for small bins of 25 singular vectors to get a more fine-grained appreciation of the evolution of the RMSE across the SVD spectrum.

Concretely, what those results suggest is that every ad-hoc coordination scenario characterized by a sampling pattern comprising more than 225, 250 and 1,400 vectors respectively will have to select singular vectors in the compatibility zone in order to minimize the RMSE. And there is every reason to expect that the order of magnitude of the number of vectors sampled during a realistic coordination scenario will be even higher than that. SimVerb, on that matter, may provide an interesting perspective, as it almost systematically leads to larger sampled sets of singular vectors: closer to 300 average, while MEN and SimLex remain at 200 (see **Table 3**). One could assume first such differences to constitute byproducts of the number of constraints encoded by each dataset: SimVerb is indeed supposed to characterize the same notion of *similarity* than SimLex but does so on a much larger sample of word pairs (3,500 vs. 999). Yet, *quantity* may not be the sole key factor here, as MEN also characterizes constraints on about 3,000 word pairs, with a similar sampling average than SimLex.

The *quality* and *nature* of those constraints may prove more determinant indeed: SimVerb encodes more fine-grained nuances on a much narrower conceptual domain in comparison to the other datasets, which could explain why it actually requires additional singular vectors to be characterized. Furthermore, we will argue here that the nature of its constraints probably makes SimVerb a much more adequate and representative lexical similarity dataset for the task at hand. Coordination, we would argue, is indeed probably better approximated by the idea that speakers align their similarity judgments on verbs like *enforce* and *impose*, rather than on the fact that *automobile* and *car* should be deemed related while *dog* and *silver* should not, as in MEN, or on the fact that *arm* and *shoulder* should be deemed similar, while *hard* and *easy* should not, as in SimLex.

If our intuition is correct, then maybe what we need in computational linguistics to better model coordination are lexical similarity datasets that encode very nuanced distinctions between lexical items, rather than broad semantic categorizations. In any case, it does not seems completely unreasonable to assume that, in a realistic ad-hoc coordination scenario, sampled vectors will ultimately fall into the compatibility zone in order to minimize the RMSE. All in all, compatibility should matter then in order to optimize structural conceptual alignment.

6.2. Compatibility Emerges From Idiosyncrasy

Considering it plausible for singular vectors to be sampled from the compatibility zone is one thing, but it does not tell us *how many of them* will actually be sampled. In order to make a point about the significance of the compatibility phenomenon, we must first indeed guarantee that the number of vectors sampled from the compatibility zone will not be marginal in comparison to the agreement zone. Is the size of the compatibility zone reported in **Section 6.1**, then, a reasonable approximation of the reality or a mere artifact of our experimental setup?

To answer this question, we must first understand where this compatibility phenomenon comes from. Recall from **Section 5.3** that the compatibility zone corresponds to the lower components of the SVD which capture more idiosyncratic information regarding the corpus at hand, in comparison to the top components which capture more systematic information about language. Agreement and compatibility are therefore first and foremost characterized by different distributional patterns across corpora, themselves deriving from differences over co-occurrence *counts*. Indeed, count-based DSMs only aggregate information from word-context co-occurrences, so that differences across aggregated distributional patterns are necessarily byproducts of cascading differences originating from the raw count matrices (recall **Figure 2**).

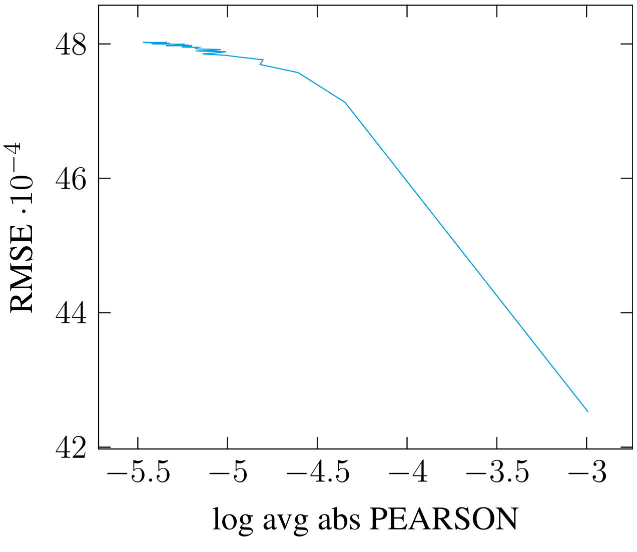


FIGURE 7 | Evolution of RMSE with log of average absolute Pearson correlation for aligned bins of 250 consecutive singular vectors sampled across [0, 10 000] on OANC and WIKI07, for DSMs with vocabularies aligned *before* the SVD step.

Yet, this particular focus on co-occurrence counts glosses over an important modeling choice of ours: in our experimental setup, DSM vocabularies are aligned *after* the SVD step, and not *before*. Therefore, the raw count matrix of a particular DSM may aggregate information over context words that are absent from other DSMs. In effect, this is tantamount to assuming that different speakers could process external stimuli from a different set of cognitive receptors, or that they could process external stimuli from a shared set of cognitive receptors but that some of those receptors will only be triggered in specific speakers.

And how much would the set of receptors differ across speakers then? Pretty much, according to our results: for the OANC–WIKI07 pair for instance, 36% of the words in OANC are not found in WIKI07, while 62% of the words in WIKI07 are not found in OANC. Note, however, that due to the Zipfian distribution of words in each corpus (Zipf, 1936; Zipf, 1949) those out-of-shared-vocabulary words only account for 2% and 3% of the total corpus word counts respectively.

What happens, then, if we align vocabularies across DSMs *before* the SVD step and filter out context columns of the original raw count matrices for words outside of the shared vocabulary? Our results, displayed in **Figure 7**, show that *the phenomenon of compatibility almost completely disappears.*

Those results have fundamental consequences for the socialization hypothesis. Indeed, they show that, if differences across speakers' background experiences are to be understood as differences in distributional patterns over external stimuli triggering a shared set of cognitive receptors, then in fact *alignment equates agreement* so that it should indeed be impossible for speakers to coordinate and align their respective conceptual spaces if those are grounded in fundamentally different background experiences.

Of course the aforementioned considerations could be deemed artifactual of the SVD and more specifically of its sensitivity to null values in the original PPMI matrix: Landauer and Dumais (1997), for instance, already noted that “a change in the value of any cell in the original matrix can, and usually does, change every coefficient in every condensed word vector” (see p. 218), while Levy and Goldberg (2014), citing (Koren et al., 2009), stressed how SVD is known to suffer from unobserved values (see p. 6). But this would only provide a technical explanation while the main question remains: should we consider this artifact to be present in human cognition as well? Probably so, at least if we are to consider conceptual knowledge to emerge from contingency-based aggregation and covariation-based decomposition of distributional information (see **Section 3.3.3**).

ll in all, our results show that *compatibility emerges from idiosyncrasy*, but that idiosyncracy here should not be understood as a distinctive difference in the *distribution* of information across background experiences, but as a difference of *nature*. Compatibility, so it seems, emerges from the *uniqueness* of each speaker and from aspects of their background experiences that uniquely distinguish them from others. Coordination, then, is enabled by what makes speakers *unique* rather than *different* from one-another.

Yet, is it completely realistic to consider that the background experience of a speaker could be primarily constituted (for more than 60% as our results above suggest) of stimulus components not experienced at all by other speakers, even if those stimuli account for a tiny portion of the overall experienced stimuli? Interestingly, those considerations directly connect us with the longstanding debate in cognitive science regarding the nature of conceptual knowledge. The fundamental question, as Huebner and Willits (2018) frame it, is really whether “knowledge consists primarily (or exclusively) of a rich sets of associations between sensory-motor features, or instead also consists of abstract, amodal concepts that bind those features together”. For if indeed conceptual knowledge is to be aggregated mostly from sensorimotor experience, it seems dubious to consider contextual vectors in DSMs to model anything but low-level core cognitive components, necessarily shared across speakers. All the more so if we are to follow previous approaches detailed in **Section 3.3.1** and consider distributional linguistic information to mirror distributional information grounded in sensorimotor experience.

But if, however, we are to consider conceptual knowledge to be aggregated mostly from *pre-existing* intermediate conceptual knowledge, a new perspective opens. Most concepts become *complex* concepts, and DSMs now model distributional learning mediated by a speaker-specific intermediate cognitive layer, rather than a set of universal core cognitive components. An unexpected solution to our puzzle appears to rest on the possible compromise between two seemingly incompatible approaches to human cognition.

7. CONCLUSION

Do speakers of the same linguistic community share similar concepts given that they are exposed to similar environments and operate in highly-coordinated social contexts? In as much as the notion of *similarity* hereby specified entails *agreement* between speakers and their conceptual spaces, the claim remains to be proven, for non-trivial conceptual variability between speakers systematically observed across experimental setups continues to be a major obstacle to be accounted for.

Yet, if we are to distinguish within similarity the notion of *agreement* from that of *compatibility*, new perspectives open: speakers no longer need to converge to *close-enough* conceptual representations in order to successfully communicate, for agreement is no longer necessary when you can merely *avoid conflict* by aligning your non-identical but nonetheless compatible representations. Even more so as this notion of compatibility leaves ample room for adjustments across speakers and thus, ultimately, successful coordination and communication. From latent compatibility to superficial agreement: all we need is a tiny conceptual shift in our characterization of similarity.

Although the cognitive plausibility of our proposed model remains to be assessed, it already provides an intuitive explanation to the very problem of conceptual variability, henceforth conceived as a mere artifact of conceptual compatibility. Indeed, our experimental approach shows that the number of compatible subspaces largely extend the number of agreeing ones, so that speakers can never be expected to agree more than to some extent. Conceptual variability should therefore not be seen as a byproduct of faulty experimental setups, but rather as a key property of human cognition.

All in all, the socialization hypothesis may very well prove to be an unnecessary prerequisite to successful communication. But our study suggests implicitly that other assumptions grouding standard models of communication could also prove unnecessary, if not unfounded. The *identicity of messages*, assumed to characterize communication success in a standard Shannon–Weaver code model, could be one of them.

All things considered indeed, communication may probably be best formalized as the cooperative act of *avoiding conflict*, rather than maximizing agreement.

DATA AVAILABILITY STATEMENT

All data and softwares used throughout this work can be found at https://gitlab.com/akb89/avoiding-conflict

AUTHOR CONTRIBUTIONS

AK came up with the original idea, designed and carried out the experiments. AH supervised the work. Both authors contributed to the writing of the paper.

ACKNOWLEDGMENTS

Authors are grateful to Ludovica Pannitto for feedback on early version of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2020.523920/full#supplementary-material.

REFERENCES

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: distributed word representations for multilingual NLP. Available from: https://arxiv.org/abs/1307.1662

Altmann, G. T. M. (1997). *The Ascent of Babel: an exploration of language, mind, and understanding*. Oxford, United Kingdom: Oxford University Press.

Anderson, R. C., and Ortony, A. (1975). On putting apples into bottles — a problem of polysemy. *Cognit. Psychol.* 7, 167–180. doi:10.1016/0010-0285(75)90008-0

Andrews, H. C., and Patterson, C. L. (1976a). Singular value decomposition (SVD) image coding. *IEEE Trans. Commun.* 24, 425–432. doi:10.1109/TCOM.1976.1093309

Andrews, H. C., and Patterson, C. L. (1976b). Singular value decompositions and digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* 24, 26–53. doi:10.1109/TASSP.1976.1162766

Aslin, R. N., and Newport, E. L. (2012). Statistical learning: from acquiring specific items to forming general rules. *Curr. Dir. Psychol. Sci.* 21, 170–176. doi:10.1177/0963721412436806

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychol. Sci.* 9, 321–324. doi:10.1111/1467-9280.00063

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.

Barker, C. (2002). The dynamics of vagueness. *Ling. Philos.* 25, 1–36. doi:10.1023/A:1014346114955

Baroni, M., Dinu, G., and Kruszewski, G. (2014). "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in Proceedings of the 52nd annual meeting of the association for computational linguistics, Baltimore, MD, June 2014 (Baltimore, MD: Association for Computational Linguistics), 238–247. doi:10.3115/v1/P14-1023

Barsalou, L. W. (2017). "Cognitively plausible theories of concept composition," in *Compositionally and concepts in linguistics and psychology*. Editors J. A. Hampton and Y. Winter (Cham, United Kingdom: Springer International Publishing), 9–30.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). "Language and simulation in conceptual processing," in *Symbols and embodiment*. Editors M. De Vega, A. M. Glenberg, and A. C. Graesser (Oxford, United Kingdom: Oxford University Press), 245–283.

Barsalou, L. W. (1987). "The instability of graded structure: implications for the nature of concepts," in *Concepts and conceptual development: ecological and intellectual factors in categorization*. Editor U. Neisser (Cambridge, UK: Cambridge University Press), 101–140.

Bird, S., Dale, R., Dorr, B., Gobson, B., Joseph, M., Kan, M.-Y., et al. (2008). "The ACL anthology reference corpus: a reference dataset for bibliographic research in computational linguistics," in Proceedings of the sixth international conference on language resources and evaluation (LREC'08), Marrakech, Morocco, May 2008 (Marrakech, Morocco: European language resources association (ELRA)), 1–5.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Brand, M. (2006). Fast low-rank modifications of the thin singular value decomposition. *Lin. Algebra Appl.* 415, 20–30. doi:10.1016/j.laa.2005.07.021

Brand, M. (2003). "Fast online SVD revisions for lightweight recommender systems," in Proceedings of the 2003 SIAM international conference on data mining, Cambridge, MA, April 2003 (Cambridge, MA: SIAM), 37–46. doi:10.1137/1.9781611972733.4

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493. doi:10.1037/0278-7393.22.6.1482

Brennan, S. E. (1998). "The grounding problem in conversations with and through computers," in *Social and cognitive approaches to interpersonal communication*. Editors S. R. Fussell and R. J. Kreuz (New York, NY: Psychology Press), 201–225. doi:10.4324/9781315805917

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.* 49, 1–47. doi:10.1613/jair.4135

Bullinaria, J. A., and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behav. Res. Methods* 39, 510–526. doi:10.3758/BF03193020

Bullinaria, J. A., and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behav. Res. Methods* 44, 890–907. doi:10.3758/s13428-011-0183-8

Bunch, J. R., and Nielsen, C. P. (1978). Updating the singular value decomposition. *Numer. Math.* 31, 111–129. doi:10.1007/BF01397471

Businger, P. A. (1970). Contribution no 26. Updating a singular value decomposition. *BIT* 10, 376–397. doi:10.1007/BF01934207

Caron, J. (2001). *Experiments with LSA scoring: optimal rank and basis*. Boulder, CO: University of Colorado.

Casasanto, D., and Lupyan, G. (2015). "All concepts are ad hoc concepts," in *The conceptual mind. New directions in the study of concepts*. Editors E. Margolis and S. Laurence (Cambridge, MA: MIT Press), 543–566.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognit. Sci.* 5, 121–152. doi:10.1207/s15516709cog0502_2

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi:10.1016/0010-0277(86)90010-7

Clark, H. H. (1992). *Arenas of language use*. Chicago, IL: University of Chicago Press.

Clark, H. H. (1983). "Making sense of nonce sense," in *The process of language understanding*. Editors G. F. d'Arcais and R. Jarvella (John Wiley & Sons), 297–331.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Clark, S. (2012). "Vector space models of lexical meaning," in *Handbook of contemporary semantics*. 2nd Edn. Editors S. Lappin and C. Fox (Wiley-Blackwell), 493–522.

Collobert, R., and Weston, J. (2008). "A unified architecture for natural language processing: deep neural networks with multitask learning," in Proceedings of the 25th international conference on machine learning, New York, NY, April 2008 (New York, NY: Association for Computing Machinery), 160–167. doi:10.1145/1390156.1390177

Connell, L, and Lynott, D. (2014). Principles of representation: why you can't represent the same concept twice. *Top Cogn Sci* 6, 390–406. doi:10.1111/tops.12097

Connell, L. (2019). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Lang. Cogn. Neurosci.* 34, 1308–1318. doi:10.1080/23273798.2018.1471512

Cooper J. M. (Editor) (1997). Plato Complete Works. Indianapolis: Hackett., (Indianapolis, IN Hackett).

Cruse, D. A. (1986). *Lexical semantics*. Cambridge, UK: Cambridge University Press.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford, UK: Oxford University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dev, S., Hassan, S., and Phillips, J. M. (2018). Absolute orientation for word embedding alignment. Available from: https://www.groundai.com/project/absolute-orientation-for-word-embedding-alignment/1

De Vega, M., Glenberg, A., and Graesser, A. (2008). *Symbols and embodiment: debates on meaning and cognition*. (Oxford, United Kingdom: Oxford University Press).

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Plunkett, K., and Parisi, D. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press.

Elman, J. L. (1990). Finding structure in time. *Cognit. Sci.* 14, 179–211. doi:10.1207/s15516709cog1402_1

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi:10.1016/0010-0277(93)90058-4

Erk, K. (2012). Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguistics Compass* 6, 635–653. doi:10.1002/lnco.362

Evans, V. (2009). *How words mean: lexical concepts, cognitive models, and meaning construction*. Oxford, UK: Oxford University Press.

Federmeier, K. D., and Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495. doi:10.1006/jmla.1999.2660

Fodor, J. D. (1977). *Semantics: theories of meaning in generative grammar*. Cambridge, MA: Harvard University Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Commun. ACM* 30, 964–971. doi:10.1145/32206.32212

Gärdenfors, P. (2004). *Conceptual spaces: the geometry of thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: semantics based on conceptual spaces*. Cambridge, MA: MIT Press.

Garrod, S., and Pickering, M. J. (2009). Joint action, interactive alignment, and dialogue. *Top. Cogn. Sci.* 1, 292–304. doi:10.1111/j.1756-8765.2009.01020.x

Gasparri, L., and Marconi, D. (2019). "Word meaning," in *The stanford encyclopedia of philosophy*. Editor E. N. Zalta (Stanford, CA: Metaphysics Research Lab, Stanford University).

Gawronski, B., and Cesario, J. (2013). Of mice and men: what animal research can tell us about context effects on automatic responses in humans. *Pers. Soc. Psychol. Rev.* 17, 187–215. doi:10.1177/1088868313480096.PMID:23470281

Gentle, J. E. (2009). "Numerical linear algebra", in *Statistics and computing* (New York, NY: Springer), 203–240.

Gentner, D. (1988). Metaphor as structure mapping: the relational shift. *Child Dev.* 59, 47–59. doi:10.2307/1130388

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). "Simverb-3500: a large-scale evaluation set of verb similarity" in Proceedings of the 2016 conference on empirical methods in natural language processing, Cambridge, United States, November 2016 (Cambridge, United States: Association for Computational Linguistics), 2173–2182. doi:10.18653/v1/D16-1235

Gleitman, L., and Papafragou, A. (2012). "New perspectives on language and thought," in *The oxford handbook of thinking and reasoning*. Editors K. J. Holyoak and R. G. Morrison (Oxford, UK: Oxford University Press), 543–568.

Glenberg, A. M., and Mehta, S. (2008). Constraint on covariation: it's not meaning. *Rivista di Linguistica (Italian Journal of Linguistics)* 20, 241–264.

Goldstone, R. L., and Son, J. Y. (2012). "Similarity," in *The oxford handbook of thinking and reasoning*. Editors K. J. Holyoak and R. G. Morrison (Oxford, UK: Oxford University Press), 155–176.

Goodman, N. (1972). "Seven strictures on similarity," in *Problems and projects*. Editor N. Goodman (New York, NY: The Bobbs-Merrill Co), 437–446.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. New York, NY: Springer.

Grice, H. P. (1975). "Logic and conversation," in *Speech Acts* (Leiden, Netherlands: Brill), 41–58. doi:10.1163/9789004368811_003

Grice, H. P. (1969). Utterer's meaning and intention. *Phil. Rev.* 78, 147–177.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). *Topics in semantic representation* 114, 211–244. doi:10.1037/0033-295X.114.2.211

Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. *Perspect. Psychol. Sci.* 14, 1006–1033. doi:10.1177/1745691619861372

Hall, J., Owen Van Horne, A., and Farmer, T. (2018). Distributional learning aids linguistic category formation in school-age children. *J. Child Lang.* 45, 717–735. doi:10.1017/S0305000917000435

Harnad, S. (1990). The symbol grounding problem. *Phys. Nonlinear Phenom.* 42, 335–346. doi:10.1016/0167-2789(90)90087-6

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi:10.1080/00437956.1954.11659520

Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., Hanke, M., and Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72, 404–416. doi:10.1016/j.neuron.2011.08.026

Heider, E. R., and Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognit. Psychol.* 3, 337–354. doi:10.1016/0010-0285(72)90011-4

Heilman, K. M., Tucker, D. M., and Valenstein, E. (1976). A case of mixed transcortical aphasia with intact naming. *Brain* 99, 415–426. doi:10.1093/brain/99.3.415

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: evaluating semantic models with genuine similarity estimation. *Comput. Ling.* 41, 665–695. doi:10.1162/COLI_a_00237

Huebner, P. A., and Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Front. Psychol.* 9, 133.

Humboldt, W. V. (1836/1988). *Character of languages*. Cambridge, UK: Cambridge University Press.

Hutchinson, J. W., and Lockhead, G. R. (1977). Similarity as distance: a structural principle for semantic memory. *J. Exp. Psychol. Hum. Learn. Mem.* 3, 660–678. doi:10.1037/0278-7393.3.6.660

Jenkins, J. J. (1954). "Transitional organization: association techniques," in *Psycholinguistics. A survey of theory and research problems*. Editors C. E. Osgood and T. A. Sebeok (Baltimore, MD: Waverly Press, Inc.), 112–118.

Jones, M. N., Willits, J., and Dennis, S. (2015). "Models of semantic memory," in *The oxford handbook of computational and mathematical psychology*. Editors J. R. Busemeyer, Z. Wang, J. R. Twnsend, and A. Eidels (Oxford, United Kingdom: Oxford University Press). doi:10.1093/oxfordhb/9780199957996.013.11

Kemmerer, D., Rudrauf, D., Manzel, K., and Tranel, D. (2012). Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. *Cortex* 48, 826–848. doi:10.1016/j.cortex.2010.11.001

Kiefer, M., Adams, S. C., and Zovko, M. (2012). Attentional sensitization of unconscious visual processing: top-down influences on masked priming. *Adv. Cognit. Psychol.* 8, 50–61. doi:10.2478/v10053-008-0102-4

Kiela, D., and Clark, S. (2014). "A systematic study of semantic vector space model parameters," in Proceedings of the 2nd workshop on continuous vector space models and their compositionality (CVSC), Cambridge, United States, April 2014 (Cambridge, MA: Association for Computational Linguistics), 21–30. doi:10.3115/v1/W14-1503

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 42, 30–37. doi:10.1109/MC.2009.263

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). "What substitutes tell us - analysis of an "all-words" lexical substitution corpus," in Proceedings of the 14th conference of the European chapter of the association for computational linguistics, Gothenburg, Sweden, April 2014. (Gothenburg, Sweden: Association for Computational Linguistics), 540–549. doi:10.3115/v1/E14-1057

Labov, W. (1973). "The boundaries of words and their meanings," in *New ways of analyzing variation in English* Editors C.-J. N. Bailey and R. W. Shuy (Washington, United States: Georgetown University Press), 340–373.

Landau, B., Dessalegn, B., and Goldberg, A. M. (2010). "Language and space: momentary interactions," in *Language, cognition and space: the state of the art and new directions*. Editors V. Evans and P. Chilton (Sheffield, United Kingdom: Equinox eBooks Publishing), 51–78. doi:10.1558/equinox.22024

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240.

Lapesa, G., and Evert, S. (2014). A large scale evaluation of distributional semantic models: parameters, interactions and model selection. *Trans. Assoc. Comput. Linguist.* 2, 531–546. doi:10.1162/tacl_a_00201

Lasersohn, P. (1999). Pragmatic halos. *Language* 75, 522–551. doi:10.2307/417059

Lebois, L. A., Wilson-Mendenhall, C. D., and Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognit. Sci.* 39, 1764–1801. doi:10.1111/cogs.12174

Lenci, A. (2018). Distributional models of word meaning. *Ann. Rev. Linguist.* 4, 151–171. doi:10.1146/annurev-linguistics-030514-125254

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. 2nd Edn. Cambridge, United States: Cambridge University Press.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput. Linguist.* 3, 211–225. doi:10.1162/tacl_a_00134

Levy, O., and Goldberg, Y. (2014). "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems 27*. Editors Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Red Hook, NY: Curran Associates, Inc.), 2177–2185.

Lewis, D. K. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Louwerse, M., and Connell, L. (2011). A taste of words: linguistic context and perceptual simulation predict the modality of words. *Cognit. Sci.* 35, 381–398. doi:10.1111/j.1551-6709.2010.01157.x

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Top Cogn Sci* 3, 273–302. doi:10.1111/j.1756-8765.2010.01106.x

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* 28, 203–208. doi:10.3758/BF03204766

Malt, B. C., Gennari, S. P., Imai, M., Ameel, E., Saji, N., and Majid, A. (2015). "Where are the concepts? What words can and can't reveal," in *The conceptual mind. New directions in the study of concepts*. Editors E. Margolis and S. Laurence (Cambridge, MA: MIT Press).

Malt, B. C., and Sloman, S. A. (2007). "Artifact categorization: the good, the bad, and the ugly," in *Creations of the mind: theories of artifacts and their representation*. Editors E. Margolis and S. Laurence (Oxford, UK: Oxford University Press), 85–123.

Malt, B. C. (2019). Words, thoughts, and brains. *Cogn. Neuropsychol.* 37, 241–253. doi:10.1080/02643294.2019.1599335

Mandera, P., Keuleers, E., and Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *J. Mem. Lang.* 92, 57–78. doi:10.1016/j.jml.2016.04.001

Mandler, J. M., Bauer, P. J., and McDonough, L. (1991). Separating the sheep from the goats: differentiating global categories. *Cognit. Psychol.* 23, 263–298. doi:10.1016/0010-0285(91)90011-C

Marconi, D. (1997). *Lexical competence*. Cambridge, MA: MIT Press.

Margolis, E., and Laurence, S. (2019). "Concepts," in *The stanford encyclopedia of philosophy*. Editor E. N. Zalta. (Stanford, CA: Metaphysics Research Lab, Stanford University).

Martin, D. I., and Berry, M. W. (2007). *Mathematical foundations behind latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCarthy, D., and Navigli, R. (2009). The English lexical substitution task. *Comput. Humanit.* 43, 139–159. doi:10.1007/s10579-009-9084-1

McCarthy, J., and Lifschitz, V. (1989). *Formalizing common sense*. Norwood, NJ: Ablex Publishing Corporation.

McDonald, S., and Ramscar, M. (2001). Testing the distributioanl hypothesis: the influence of context on judgments of semantic similarity. Proceedings of the annual meeting of the cognitive science society Edinburgh, Scotland, August 1-4, 2001.

McRae, K., and Jones, M. N. (2013). "Semantic memory," in *The oxford handbook of cognitive psychology*. Editor D. Reisberg (Oxford, UK: Oxford University Press). doi:10.1093/oxfordhb/9780195376746.013.0014

Medin, D., Goldstone, R., and Gentner, D. (1993). Respects for similarity. *Psychol. Rev.* 100, 254–278. doi:10.1037/0033-295X.100.2.254

Melara, R. D., Marks, L. E., and Lesko, K. E. (1992). Optional processes in similarity judgments. *Percept. Psychophys.* 51, 123–133. doi:10.3758/BF03212237

Merriman, W. E., Schuster, J. M., and Hager, L. (1991). Are names ever mapped onto preexisting categories? *J. Exp. Psychol. Gen.* 120, 288–300. doi:10.1037//0096-3445.120.3.288

Mervis, C. B. (1987). "Child-basic object categories and early lexical development," in *Concepts and conceptual development: ecological and intellectual factors in categorization*. Editor U. Neisser (Cambridge, UK: Cambridge University Press), 233.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). "Distributed representations of words and phrases and their compositionality," in Proceedings of the 26th international conference on neural information processing systems, Red Hook, NY, October 2013 (Red Hook, NY: Curran Associates Inc.), 3111–3119.

Mu, J., and Viswanath, P. (2018). "All-but-the-Top: simple and effective postprocessing for word representations," in International conference on learning representations, Vancouver, BC, Canada, February 2018. (Vancouver, BC, Canada: ICLR), 1–25.

Murphy, G., and Andrew, J. (1993). The conceptual basis of antonymy and synonymy in adjectives. *J. Mem. Lang.* 32, 301–319. doi:10.1006/jmla.1993.1016

Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289–316. doi:10.1037/0033-295X.92.3.289

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychol. Bull.* 49, 197–237. doi:10.1037/h0055737

Pandey, A. K., and Heilman, K. M. (2014). Conduction aphasia with intact visual object naming. *Cognit. Behav. Neurol.* 27, 96–101. doi:10.1097/WNN.0000000000000029

Papini, M. R., and Bitterman, M. E. (1990). The role of contingency in classical conditioning. *Psychol. Rev.* 97, 396–403. doi:10.1037/0033-295x.97.3.396

Pelletier, F. J. (2017). "Compositionally and concepts—a perspective from formal semantics and philosophy of language," in *Compositionally and concepts in linguistics and psychology*. Editors J. A. Hampton and Y. Winter (Cham, UK: Springer International Publishing), 31–94. doi:10.1007/978-3-319-45977-6_3

Pickering, M. J., and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347. doi:10.1017/S0140525X12001495

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–226. doi:10.1017/S0140525X04000056

Pickering, M. J., and Garrod, S. (2006). Alignment as the basis for successful communication. *Res. Lang. Comput.* 4, 203–228. doi:10.1007/s11168-006-9004-0

Putnam, H. (1975). "The meaning of 'meaning'," in *Complete works*. (Cambridge, UK: Cambridge University Press), 215–271. doi:10.1017/CBO9780511625251.014

Rastogi, P., Van Durme, B., and Arora, R. (2015). "Multiview LSA: representation learning via generalized CCA," in Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies, Denver, CO, May 2015 (Denver, CO: Association for Computational Linguistics), 556–566. doi:10.3115/v1/N15-1058

Raunak, V., Gupta, V., and Metze, F. (2019). "Effective dimensionality reduction for word embeddings," in Proceedings of the 4th workshop on representation learning for NLP (RepL4NLP-2019), Florence, Italy, August 2019 (Florence, Italy: Association for Computational Linguistics), 235–243. doi:10.18653/v1/W19-4328

Recanati, F. (2004). *Literal meaning*. Cambridge, UK: Cambridge University Press.

Redington, M., Crater, N., and Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognit. Sci.* 22, 425–469. doi:10.1016/S0364-0213(99)80046-9

Rescorla, M. (2019). "Convention," in *The stanford encyclopedia of philosophy*. Editor E. N. Zalta. (Stanford, CA: Metaphysics Research Lab, Stanford University).

Rescorla, R. A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. *J. Comp. Physiol. Psychol.* 66, 1–5. doi:10.1037/h0025984

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations on the effectiveness of reinforcement and non-reinforcement," in *Classical conditioning II: current research and theory*. Editors A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rogers, T. T., and McClelland, J. L. (2004). *Semantic cognition: a parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* 104, 192–233.

Rosch, E. (1973). "On the internal structure of perceptual and semantic categories," in *Cognitive development and acquisition of language*. Editor T. E. Moore (San Diego, CA: Academic Press), 111–144. doi:10.1016/B978-0-12-505850-6.50010-4

Rosch, E. (1978). "Principles of categorization," in *Cognition and categorization*. Editors E. Rosch and B. Lloyd (Hillsdale, NJ: Lawrence Elbaum Associates), 27–48.

Ross, B. H., and Murphy, G. L. (1999). Food for thought: cross-classification and category organization in a complex real-world domain. *Cognit. Psychol.* 38, 495–553. doi:10.1006/cogp.1998.0712

Roth, E. M., and Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognit. Psychol.* 15, 346–378. doi:10.1016/0010-0285(83)90012-9

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Curr. Dir. Psychol. Sci.* 12, 110–114. doi:10.1111/1467-8721.01243

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)* 20, 33–53.

Schiffer, S. R. (1972). *Meaning*. Oxford, United Kingdom: Oxford University Press.

Schütze, H. (1992). "Dimensions of meaning," in Proceedings of the 1992 ACM/IEEE conference on supercomputing, Los Alamitos, CA, November 1992. (Los Alamitos, CA: IEEE Computer Society Press), 787–796.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi:10.1017/S0140525X00005756

Shannon, C. E., and Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. ii. *Psychometrika* 27, 219–246. doi:10.1007/BF02289621

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika* 27, 125–140. doi:10.1007/BF02289630

Sinha, R., and Mihalcea, R. (2014). Explorations in lexical sample and all-words lexical substitution. *Nat. Lang. Eng.* 20, 99–129. doi:10.1017/S1351324912000265

Sloman, S., and Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Lang. Cognit. Process.* 18, 563–582. doi:10.1080/01690960344000035

Smith, L, and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106, 1558–1568. doi:10.1016/j.cognition.2007.06.010

Solomon, K. O., and Barsalou, L. W. (2004). Perceptual simulation in property verification. *Mem. Cognit.* 32, 244–259. doi:10.3758/BF03196856

Stalnaker, R. (2002). Common ground. *Ling. Philos.* 25, 701–721.

Stalnaker, R. (2014). *Context*. Oxford, United Kingdom: Oxford University Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 245–251. doi:10.1037/0033-2909.87.2.245

Strang, G. (2016). *Introduction to linear algebra*. 5th Edn. Cambridge, United Kingdom: Wellesley-Cambridge Press.

Tan, L., Zhang, H., Clarke, C., and Smucker, M. (2015). "Lexical comparison between Wikipedia and twitter corpora by using word embeddings," in Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, Beijing, China, July 2015 (Beijing, China: Association for Computational Linguistics), 657–661. doi:10.3115/v1/P15-2108

Tulving, E. (1972). "Episodic and semantic memory," in *Organization of memory*. Editors E. Tulving and W. Donaldson (London, UK: Academic Press), 381–402.

Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188.

Wachsmuth, I., de Ruiter, J., Jaecks, P., and Kopp, S. (2013). *Alignment in communication*. Amsterdam, Netherlands: John Benjamins Publishing Company.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Q. J. Exp. Psychol.* 27, 635–657. doi:10.1080/14640747508400525

Westera, M., and Boleda, G. (2019). "Don't blame distributional semantics if it can't do entailment," in Proceedings of the 13th international conference on computational semantics - long papers, Gothenburg, Sweden, May 2019 (Gothenburg, Sweden: Association for Computational Linguistics), 120–133. doi:10.18653/v1/W19-0410

Wilson, D., and Carston, R. (2007). "A unitary approach to lexical pragmatics: relevance, inference and ad hoc concepts," in *Pragmatics*. Editor N. Burton-Roberts (London, UK: Palgrave-Macmillan), 3.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1195–1200. doi:10.1073/pnas.1814092116

Wingfield, C., and Connell, L. (2019). Understanding the role of linguistic distributional knowledge in cognition. doi:10.31234/osf.io/hpm4z

Wolff, P., and Malt, B. C. (2010). "The language–thought interface: an introduction," in *Words and the mind*. (New York, NY: Oxford University Press). doi:10.1093/acprof:oso/9780195311129.003.0001

Yee, E., and Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychon. Bull. Rev.* 23, 1015–1027. doi:10.3758/s13423-015-0948-7

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

Zipf, G. K. (1936). *The psychobiology of language*. London, United Kingdom: Routledge.

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162–185. doi:10.1037/0033-2909.123.2.162

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 15 April 2021
doi: 10.3389/frai.2021.644554

Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in Frisian

Jelske Dijkstra[1,2]*, *Wilbert Heeringa*[1], *Lysbeth Jongbloed-Faber*[1,3] *and* *Hans Van de Velde*[1,4]

[1] *Fryske Akademy, Leeuwarden, Netherlands,* [2] *Mercator European Research Centre on Multilingualism and Language Learning, Leeuwarden, Netherlands,* [3] *Faculty of Arts and Social Sciences, Maastricht University, Maastricht, Netherlands,* [4] *Department of Languages, Literature and Communication, Utrecht University, Utrecht, Netherlands*

This paper investigates the usability of Twitter as a resource for the study of language change in progress in low-resource languages. It is a panel study of a vigorous change in progress, the loss of final t in four relative pronouns (*dy't*, *dêr't*, *wêr't*, *wa't*) in Frisian, a language spoken by ± 450,000 speakers in the north-west of the Netherlands. This paper deals with the issues encountered in retrieving and analyzing tweets in low-resource languages, in the analysis of low-frequency variables, and in gathering background information on Twitterers. In this panel study we were able to identify and track 159 individual Twitterers, whose Frisian (and Dutch) tweets posted in the era 2010–2019 were collected. Nevertheless, a solid analysis of the sociolinguistic factors in this language change in progress was hampered by unequal age distributions among the Twitterers, the fact that the youngest birth cohorts have given up Twitter almost completely after 2014 and that the variables have a low frequency and are unequally spread over Twitterers.

Keywords: CMC, Frisian, relative pronoun, t-deletion, panel study, frequency, methodology

Edited by:
James N. Stanford,
Dartmouth College, United States

Reviewed by:
Mihaela Vela,
Saarland University, Germany
Chaya Liebeskind,
Jerusalem College of Technology, Israel

Correspondence:
Jelske Dijkstra
jdijkstra@fryske-akademy.nl

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *21 December 2020*
Accepted: *09 March 2021*
Published: *15 April 2021*

Citation:
Dijkstra J, Heeringa W, Jongbloed-Faber L and Van de Velde H (2021) Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in Frisian. Front. Artif. Intell. 4:644554. doi: 10.3389/frai.2021.644554

INTRODUCTION

Since the spread of the Internet and social media, language use on the Internet has drawn the attention of scholars in linguistics (Herring, 1996; Crystal, 2001) and communication (Thurlow et al., 2004). It resulted in numerous studies on various topics within the domain of computer-mediated communication (CMC) such as: bilingual practices (Cunliffe et al., 2013; Androutsopoulos, 2014; Jongbloed-Faber et al., 2016, 2017; Reershemius, 2017; Cutler and Røyneland, 2018), discourse strategies (Herring, 2001; Herring and Paolillo, 2006; Baron, 2010; Androutsopoulos, 2014), and spelling skills (Plester et al., 2008; Stæhr, 2015). Scholars studying language change in progress got interested in CMC too, although the number of studies remains relatively low. For example, Eisenstein (2013, 2015) showed that in American English tweets t/d-deletion depends on its phonological context, but the effect is less outspoken than in speech. Grieve et al. (2019) compared regional lexical variation in British English between Twitter data and traditional survey data. In both resources similar lexical patterns were identified, but some regional patterns showed up more clearly in Twitter than in survey data. Vandekerckhove (2006), Vandekerckhove and Nobels (2010), De Decker (2014), Grondelaers et al. (2017) and Verheijen (2017) successfully used CMC corpora to study language variation and change in Dutch, a medium-sized language and the dominant language in the written domain in the Frisian language area, the geographical focus of our study (see below).

The scarceness of variation research in CMC is partly due to predominantly anonymous contributions in CMC. Consequently, information about the writer's demographic background such as gender, age, education, birth place, hometown and social class, is not directly available (Herring, 2001; Grieve et al., 2019), which hampers a variationist sociolinguistic analysis. Participants' gender and age can often be deducted from screen names or profile descriptions and pictures, but it is time consuming to search for this information. Although computer models have been built to automatically predict age and gender for large and medium-sized languages as English and Dutch (Nguyen et al., 2013) or for multilingual data (Wang et al., 2019), such models are not available for smaller languages or dialects. Furthermore, the demographic background of users of specific social media platforms differs from the offline population's background. E.g., Twitterers in the UK and the US are more likely to be younger, better educated, students or employed, single and wealthier compared to the other Internet users and the offline population (Blank, 2016). This creates a bias in research results based on Twitter data. Blank therefore discourages the use of Twitter data in social sciences.

The stylistic characteristics of language practices has been a central topic from the early start of CMC research. The use of non-standard spelling and the presence of spoken language features show up as core linguistic features of CMC. Consequently, CMC language does not correspond to traditional written communication, in which writers generally conform to spelling rules and discard spoken features. Nevertheless, language in CMC cannot be gathered completely under the concept of oral communication either, because many contextual cues that are available in spoken language, e.g., intonation and facial expression, are not possible in CMC. In other words, language practices in CMC hold somewhere between written and oral communication (Androutsopoulos, 2014). That said, one may wonder whether CMC language is an appropriate source for language variation research. Stæhr (2015) argued that precisely the presence of colloquial features in CMC language pleads for inclusion of language from digital media in the study of language variation. De Decker et al. (2016) concluded that CMC language such as chatspeak can be a useful source to study language variation, if the variables are well-chosen and analyzed, according to the standards in variationist sociolinguistic research. Additionally, Grondelaers et al. (2017) demonstrated that, despite its limitation in number of characters, tweets are a rich resource to study morphosyntactic variation as well.

Bleaman (2020) pointed out that most of the sociolinguistic studies of social media focus on a handful of languages and that minority language are neglected, due to scarcity of data and the lack of computational tools to collect and analyze data of these languages and language varieties. He observes that CMC studies of low-resource varieties have mainly focused on macro-level analyses, and not on the analysis of linguistic variables at the micro-level. Bleaman was able to trace and analyze a syntactic change in progress in a real time corpus (2012–2019) of discussion forums written in Hassidic Yiddish, a low-resource language.

In this paper, we will further explore the possibilities that social media offer for the study of language variation and change in low-resource languages. We first give a brief introduction to Frisian and the language situation in Fryslân. More detailed information can be found in Munske et al. (2001) and Jonkman and Versloot (2018). Frisian is an autochtonous minority language spoken in the province of Fryslân (the Netherlands), where it is recognized as an official language, in addition to Dutch. Both are closely-related West-Germanic languages, but mutual intelligible. Eighty-nine percent of the inhabitants in Fryslân report to understand Frisian, whereas 69% is able to speak it (very) well. Sixty-one percent of the population, about 400,000 people, is a native speaker of Frisian (Klinkenberg et al., 2018). All speakers of Frisian speak Dutch too, and most of them write mainly in Dutch.

Although Frisian has an official written standard, most of the native speakers do not read or write Frisian. In the most recent language survey, 18% of the Frisians indicate they can write it well or very well (Klinkenberg et al., 2018), but it should be noted that this increase (doubled in comparison with surveys in the 1980s and 1990s), is linked to an increasing use of Frisian in social media. Up to today, the two regional daily newspapers are written in Dutch and only occasionally use Frisian in for example quotes (Gorter, 2001). Consequently, for most Frisians writing and reading is not an everyday activity. However, social media have changed this. Jongbloed-Faber et al. (2016) showed that 87% of Frisian-speaking teenagers use Frisian on social media to some extent. On Twitter, 29% of the Frisian-speaking adolescents indicated they use Frisian often or all the time in addressed and 24% in general tweets. On WhatsApp and in chat messages on Facebook the proportion of teenagers using Frisian is even higher. While writing skills appear the most important predictive variable for the adults' use of Frisian on social media (Jongbloed-Faber, 2015), the language use with peers and attitudes are more important than writing skills among adolescents (Jongbloed-Faber et al., 2016).

The change under investigation is the substitution of t-full relative pronouns, i.e., *dy't* 'who/that', *dêr't* 'where', *wêr't* 'where,' and *wa't* 'who(m)', by their t-less counterparts, i.e., *dy*, *dêr*, *wêr,* and *wa* respectively. This change in progress has been found in an earlier real-time study on this substitution in scripted and unscripted broadcast speech (Dijkstra et al., 2017, 2018, 2019), see Section Method for an overview of the results. In this study we attempt to get more insight in this vigorous change in progress, by analyzing a large data set. The variable is sensitive to normative grammatical rules: reference grammars prescribe the use of t-full relative pronouns (Popkema, 2018, pp. 175–177).

For several reasons we opted for Twitter as a source for our study:

- More monitoring of the writing process in comparison with WhatsApp messages (Verheijen, 2018).

- The length of the messages allows for the occurrence of more complex structures, including relative clauses (needed for our linguistic variable).
- Lack of other written media: there are almost no popular blogs written in Frisian.
- The medium is frequently used by a wide range of individuals and non-professional writers.
- As a public medium the data should be easily accessible for research.
- It exists long enough to cover a period of 10 years, a minimum needed to study language change in progress.
- Frisian teenagers reported to use Frisian frequently.
- Existence of a dataset that could help in the collection of a new dataset and enabling us to follow individuals over time.
- Allowing to study the interaction of the factors time and age.

TABLE 1 | Examples of the Frisian relative pronouns *dy't, dêr't*, *wêr't,* and *wa't*.

Relative pronoun	Example			
dy't	*de man*	*dy't*	*in boek*	*lêst*
	the man	who-REL	a book	read-3SG
	'the man who reads a book'			
dêr't	*de stêd*	*dêr't*	*er no*	*wennet*
	the city	there-REL	he now	live-3SG
	'the city he currently lives in'			
wêr't	*ik wit net*	*wêr't*	*de kaai*	*is*
	I know-1SG not	where-REL	the key	be-3SG
	'I do not know where the key is'			
wa't	*wa't*	*dat dien hat*	*is*	*in held*
	who-REL	that do-PP have-AUX	be-3SG	a hero
	'whomever has done that, is a hero'			

To sum up, the current study has three research aims:

1. Explore the issues in gathering a Twitter corpus of a low-resource language such as Frisian.
2. Get insight in the validity of Twitter data for the study of language change in progress in low-resource languages.
3. Refine existing sociolinguistic insights in a vigorous change in progress in Frisian relative pronouns.

Real-Time Change in Frisian Relative Pronouns

The Frisian relative pronouns ending in'*t* are *dy't* 'who/that', *dêr't* 'where', *wêr't* 'where', and *wa't* 'who(m)'. Examples of these pronouns are shown in **Table 1**.

The relative pronoun *dy't* 'who/that' is used with feminine, masculine and plural antecedents. There are two Frisian relative adverbs, namely *dêr't* and *wêr't* 'where'. When the relative clause has an antecedent that is a location, then the adverb *dêr't* is used. In free relatives, *wêr't* is used. Due to the influence of Dutch, *dêr't* is often substituted by *wêr't* (De Haan, 2001; Dijkstra et al., 2018; Taalportaal|Relative pronouns, 2018) since the Dutch equivalent for both is *waar* 'where' which translates directly to *wêr* in Frisian (the Dutch translation of *dêr* is *daar* 'there'). The relative pronoun *wa't* 'who(m)' is used in free relatives and refers to a person (Taalportaal|Relative pronouns, 2018).

The orthographic *'t* is found in other Frisian conjunctions as well. It marks the beginning of a subordinate sentence. The addition of *'t* to conjunctions is a relatively new phenomenon in Frisian. It is mentioned for the first time in *dy't* and *hwa't* [former spelling of *wa't*] as an option next to *dy* and *hwa* in a descriptive grammar of Frisian from 1889 (Van Blom, 1889 in Van der Woude, 1960).

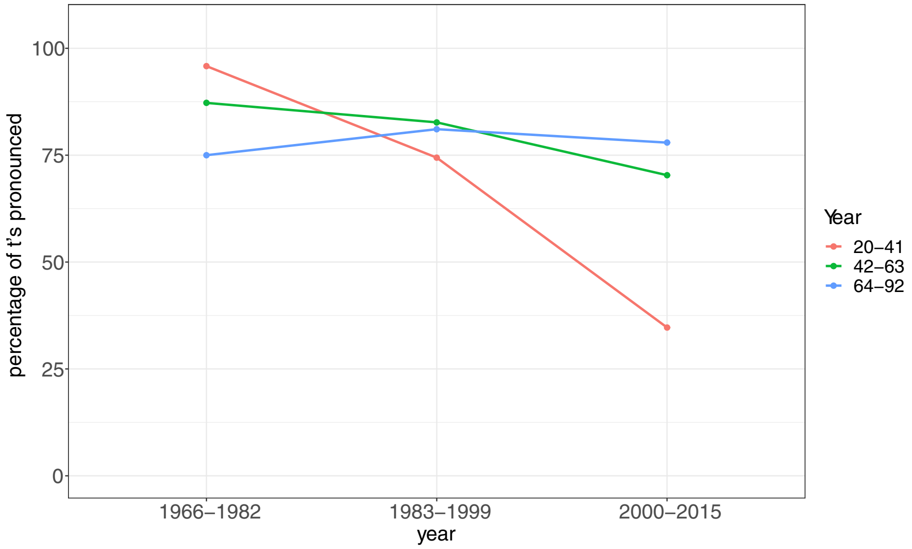


FIGURE 1 | Percentages of t-full forms in (*dy't*), (*dêr't*), (*wêr't*), (*wa't*), interaction effect between age of all speakers and broadcasting year ($n = 776$, $N = 266$). Based on: Dijkstra et al. (2019, p. 95).

Recently three real time studies on the substitution of t-full Frisian relative pronouns by t-less forms were conducted using speech data from the radio archive of the regional broadcaster *Omrop Fryslân* (Broadcasting Corporation Fryslân) Dijkstra et al. (2017, 2018, 2019) looked at this substitution in Frisian relative pronouns in non-scripted speech (i.e., spontaneous speech), semi-scripted speech (i.e., mixture of pre-written catch words/phrases and spontaneous speech) and scripted speech (all text is read aloud) (see also Chignell, 2009). The general conclusion is that the t-full forms occurred more frequently in older broadcasts and in scripted speech. The substitutions by t-less forms showed a significant increase in recent broadcasts, especially in non-scripted speech. An interaction effect between age and broadcasting year was also found (see **Figure 1**). This figure clearly shows that in the older broadcasts the youngest age group lead the change toward more t-full forms (as mentioned in the descriptive grammar from 1889 for the first time). They produce more t-full forms compared to the oldest age group in the broadcasts from 1966 to 1982. This is in line with Brouwer (1959) who observed that the realization of (t) in Frisian conjunctions was increasing. In the most recent radio fragments (2000–2015), however, we see a reversal of the first observed language change: the speakers from the oldest age group use the t-full forms most frequently, whereas the youngest age group use the t-less forms the most. This latter pattern was already observed by Van der Meer (1991) and confirmed by these studies. Finally, it should be noted that the change is now in the quick or steep phase of the traditional S-curve pattern observed in language change.

In the current study we investigate whether we can refine our insights in this language change in progress on the basis of an analysis of Twitter data, covering a time span of 10 years. By following Twitterers over time (panel study) we want to get more insight in patterns of individual stability and change during the rapid spread of a change through the community, a core theoretical topic in variationist sociolinguistics (Wagner and Buchstaller, 2017). We expect to see a continuation of the reversal previously observed in speech data, thus that the Twitterers use more t-less forms over time in the 2010's. As previously stated, language practices on Twitter are situated between oral and written language use (Androutsopoulos, 2014) and spoken language features in tweets are more common amongst young Twitters (Androutsopoulos, 2006). This means that we expect that younger Twitterers use more t-less forms than older Twitterers.

METHOD

Collecting Twitter Data

Frisian tweets have been collected earlier in the Twidentity project using a language detector trained on identifying Frisian, Limburgish and Dutch tweets (Jongbloed-Faber et al., 2017). That Twitter data set comprises 76,757 predominantly Frisian tweets of 253 Twitter accounts posted in 2013 and 2014. This list of 253 Twitter accounts consists of 208 individual Twitter accounts (71,835 predominantly Frisian tweets) and 45 Twitter accounts that were owned by SMEs or organizations. Since we intended to conduct a panel study of language change in progress in Frisian relative pronouns, we decided to build a new corpus with all tweets of the abovementioned 208 individual Twitter accounts from the Twidentity project posted from January 1, 2010 until December 31, 2019, covering a time span of 10 years in real time.

Method of Retrieving Tweets

Twitter's REST and stream Application Program Interfaces (API) are meant to be used for retrieving tweets. The R package rtweets provides several functions that use these APIs. The simplest is to use the function search_tweets, but this function only returns tweets from the past 6–9 days. Since we aim to retrieve tweets of a period of 10 years, we had rather to use the function search_fullarchive which uses Twitter's premium APIs. In order to be able to use this function, one needs to have a Twitter developer account which can be obtained for free. By using this function, we were able to retrieve 1,717 tweets after which we got the message: "Request exceeds account's current package request limits. Please upgrade your package and retry or contact Twitter about enterprise access." Therefore, we applied for an enterprise API access at https://developer.twitter.com/en/enterprise-application twice, but never received any response.

Therefore, we used GetOldTweets3, a "Python 3 library and a corresponding command line utility" developed by Jefferson Henrique and forked by Dmitry Mottl (see: https://pypi.org/project/GetOldTweets3/). The authors describe the methodology they implemented as follows:

> "Basically when you enter on a Twitter page a scroll loader starts. If you scroll down you start to get more and more tweets, all through calls to a JSON provider. After mimic we get the best advantage of Twitter Search on browsers, it can search the deepest oldest tweets."

The tweets were collected late January and early February 2020. We wanted to limit our study to tweets sent by people who are still living in the province of Fryslân and are likely to interact in Frisian on a daily basis. This was operationalized by retrieving those accounts that were registered within a radius of 50 km from the village Grou, which is centrally located in the province. Due to closure of twitter accounts and emigration from Fryslân between 2013 and 2020, the number of Twitterers in our data set is 186.

We ran the GetOldTweets3 script in order to retrieve tweets between 1 January 2010 and 31 December 2019. When doing this a second time, we obtained a set of tweets that slightly differed from the first set. Therefore, we ran the script 14 times, and combined the 14 sets of tweets into one set. Tweets that appeared multiple times were kept only once. In this way, we maximized the number of tweets that we retrieved. In total we retrieved 698,369 tweets of 186 Twitterers.

The Entire Twitter Data Set (698,369 Tweets of 186 Twitterers)

The variation between Twitterers in the production of tweets is huge, ranging from 41 to 40,887 tweets per person. The 186 Twitterers wrote on average 3,755 tweets (sd = 5,506). **Figure 2** shows the distribution of the number of tweets per Twitterer. This unequal distribution of tweets over speakers might have an impact on the distribution of our variable too, see below. But language variationists usually deal with this problem by taking per individual a sample of the variable.

Figure 3 gives the frequency distribution of the tweets in the past decade. 77.3% of the tweets were retrieved between 2010 and 2013, with a peak in 2012. From 2014 onwards the number of tweets remains stable. It should be noted that this frequency distribution is biased by our retrieval method. We only follow Twitterers that are part of the Twidentity data set, which were active in 2013 and 2014. People who started using Twitter later were not included. Twenty Twitterers posted at least 20 tweets per year. So, the frequency distribution presents the number of tweets in this panel study, it is not a representative frequency distribution of all tweets in this decade.

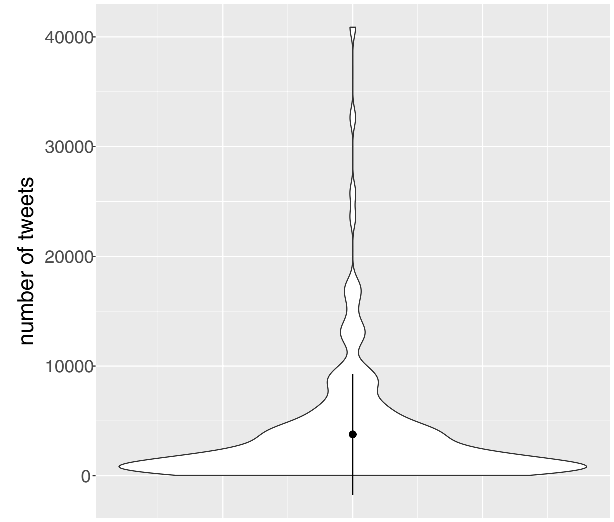


FIGURE 2 | Violin plot showing the distribution of the number of tweets per Twitterer in the period from January 1, 2010 until December 31, 2019 (n = 698,369, N = 186). The black dot represents the mean, and the vertical line represents 1 standard deviation on either side of the dot.

Figure 4 illustrates the mean number of words per tweet (averaged over Twitterers), split up by year. The mean number of words per tweet remains quite constant over the years. Twitter doubles the number of characters from 140 to 280 in 2017. As becomes clear from **Figure 4**, this has a moderate impact on the length of tweets. The tweets posted in 2018 and 2019 are a couple of words longer, however, they also show more variance in number of words. This is in line with findings of Glicoric et al. (2020) on number of characters of tweets. Before the switch to 280 characters, 9% of English tweets were exactly the maximum 140 characters long. After the switch still a substantial number of tweets reached the new maximum of 280 characters. They also demonstrated a significant but moderate increase in number of characters, also across languages.

SELECTING TWEETS WITH FRISIAN RELATIVE PRONOUNS

Data Cleaning and Coding

Once the data was retrieved, we first wanted to get more insight in the characteristics of the Twitterers. We tried to detect birth year and gender of the 186 individual Twitterers in the data set (698,369 tweets) by extensive searches on the internet and other public resources. For 82.5% of them, birth year could be retrieved. For the remaining 17.5%, the birth year was estimated based on the user's profile picture and additional public information on Facebook or LinkedIn. One of the Twitterers was an outlier, being born in 1929, since the second oldest person was born in 1945, so his tweets were discarded from the corpus.

Next, we automatically selected all tweets that contained one or more words that were similarly written as one of the target variables (see **Table 2** for all possible variants of the target

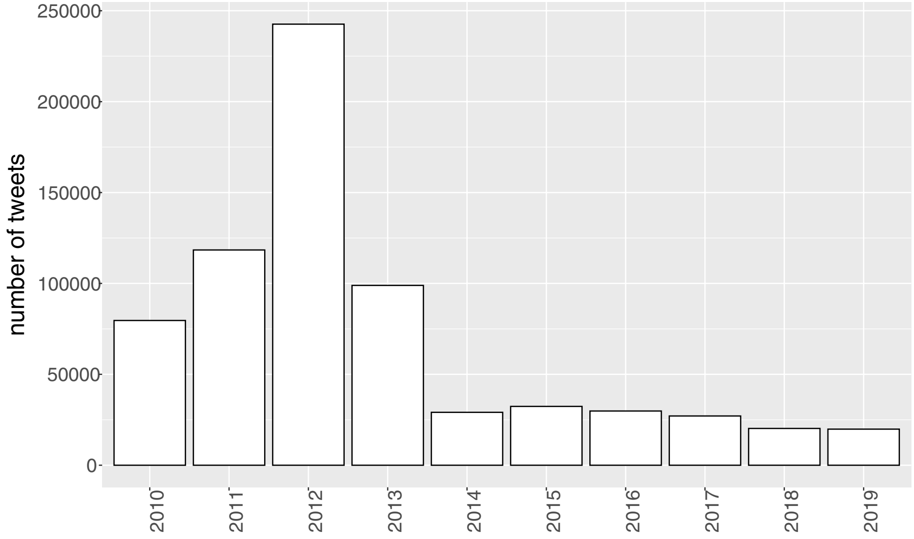


FIGURE 3 | Distribution of the number of tweets per year (n = 698,369, N = 186).

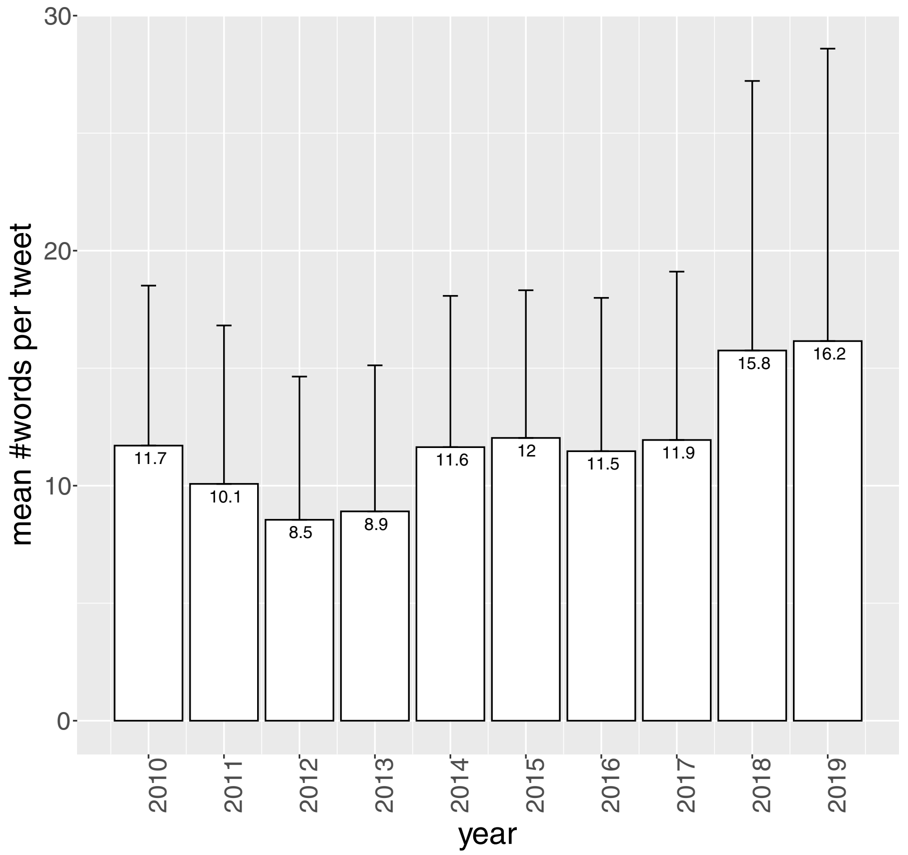


FIGURE 4 | Mean number of words per tweet (averaged over Twitterers), split up by year (n = 698,369, N = 186).

TABLE 2 | All variants of the target variables (*dy’t*), (*dêr’t*), (*wêr’t*), and (*wa’t*).

Variable	Variants
(*dy’t*)	*dy’t, dy’tst, die’t, die’tst dy, dy’st, die, die’st*
(*dêr’t*)	*dêr’t, dêr’tst, der’t, der’tst dêr, dêr’st, der, der’st*
(*wêr’t*)	*wêr’t, wêr’tst, wer’t, wer’tst wêr, wêr’st, wer, wer’st*
(*wa’t*)	*wa’t, wa’tst, wie’t, wie’tst wa, wa’st, wie, wie’st*

TABLE 3 | Results for CLD3 and textcat (n = 698,369).

CLD3	textcat	Number of tweets	%	Language
Frisian	Frisian	147,585	21.1	Frisian
Frisian	?	23,141	3.3	Frisian
?	Frisian	59,357	8.5	Frisian
Frisian	Dutch	4,530	0.6	Frisian
Dutch	Frisian	3,028	0.4	Frisian
Dutch	Dutch	141,551	20.3	Dutch
Dutch	?	18,780	2.7	Undetermined
?	Dutch	104,870	15.0	Undetermined
?	?	195,527	28.0	Undetermined

‘?’ means that the language detected was neither Frisian nor Dutch or that the detector was not able to determine the language.

variables). Note that some of these variants are also Dutch words, so this data set comprised Frisian and Dutch tweets, and tweets of which the language was undetermined. As **Table 3** shows, we also selected tweets in which the target variables had a suffix—*st* to the pronoun. This is the inflection marker for second person singular that is used when the relative clause has a second person singular subject, e.g., *de auto dêr’tst yn rydst* ‘the car which-you drive’. The (t) before the –*st* suffix is usually not pronounced (Hoekstra, 1985). The automatic selection resulted in a subset of 100,365 tweets from 185 Twitter accounts.

The variants without /t/ could also be other parts of speech, e.g., the adverb *dêr*, the verb *die* (first and third person singular past tense of *dwaan* ‘to do’), the verb *wie* (first and third person singular past tense of *wêze* ‘to be’), etc. As there was no POS-tagger for Frisian at the time of research, as for most low-resourced languages, the 100,395 tweets with possible realizations of the variables had to be checked manually by the first author who is a native speaker of Frisian. This resulted in a data set with tweets that had at least one of the relative pronouns comprised 5,500 tweets and 5,688 tokens of 159 Twitterers. During the analysis, we saw unexpected variants of the target variables where instead of <*’t*> the adverb *as, at,* **os* or **ot* was used, i.e., *wer as*, or *wer os*. Tweets with these variants were discarded. Consequently, the final data set contains 5,395 tweets and 5,559 tokens of 159 Twitter accounts.

Finally, the Twitterers were split into two groups based upon the spelling used in their tweets: a spelling following grammatical rules of Frisian and a phonetically oriented one. As mentioned, only 18% of the population of Fryslân claims to be able to write (well) in Frisian (Klinkenberg et al., 2018). Therefore, the spelling habits of the Twitterers might be a factor in the use of t-full or t-less forms. When the spelling was phonetic and/or according to Dutch spelling rules or with many Dutch interferences, the spelling was considered as phonetically oriented spelling. Otherwise, the spelling was coded as standard.

Automatic Detection of Frisian Tweets

There are three main types of tweets:

1. tweets written in Frisian following Frisian spelling rules.
2. tweets written in Frisian following Dutch spelling rules.
3. tweets written in Dutch following Dutch spelling rules.

Considering the second type, many Twitterers do not use typical Frisian characters such a <û>, <ú> or <y>, but use Dutch <oe>, <uu> or <i> instead. This type of tweets may easily be classified as Dutch by the language detectors. We wanted to see to what extent it was possible to detect automatically the Frisian and Dutch tweets. To this end we used the two language detectors that are available in the R programming language: the function textcat from the textcat package (Hornik et al., 2013), and the function detect_language from the cld3 package (Ooms, 2020). We used the following procedure:

- If one of the detectors classifies a tweet as Frisian, the tweet is coded as Frisian.
- If both detectors classify a tweet as Dutch, the tweet is coded as Dutch.
- In all other cases the language remained undetermined.

The function textcat provides an implementation of the Cavnar and Trenkle (1994) approach to text categorization based on character *n*-gram frequencies. This approach uses two steps. First, training corpora are collected for a set of languages. Each corpus includes texts all written in the same language. For each corpus the frequency distribution of all *n*-grams ($n = 1.5$) found in the texts are computed. Then the *n*-grams are sorted from the most to the least frequent. The *k* most frequent ones are retained and represent a language profile.

In the second step the language of a given text document is identified. A profile is computed, using the same procedure as for the training corpora. Then the text document is classified according to the language of the language profile with the smallest distance to the text document profile. Cavnar and Trenkle (1994) suggest the so-called "out-of-place" distance measure. When measuring the distance between the text document profile and a language profile, for each *n*-gram in the text document profile, we find its counterpart in the language profile and calculate how far out of place it is. For example, if an *n*-gram ranks the second in the text document profile, and the nineth in the language profile, the out of place is seven. If an *n*-gram is not in the category profile, it takes some maximum out-of-place value. The distance between the two profiles is the sum of all of the out-of-place values for all *n*-grams.

The function detect_language is a wrapper for Google's Compact Language Detector 3 (CLD3). CLD3 is a neural network model for language identification. Character *n*-grams are extracted from the input text and the fraction of times each of them appears is computed. For example, 'banana' has unigrams 'b', 'a,' and 'n' with fractions 1/6, 3/6, and 2/6, bigrams 'ba', 'na,' and 'an' with fractions 1/5, 2/5, and 2/5, and trigrams 'ban', 'ana,' and 'nan' with fractions 1/4, 2/4, and 1/4. This information is passed to a trained neural network model which subsequently predicts the language. See also: https://github.com/google/cld3#readme

The results for the two language detectors are presented in **Table 3**. In total, of the almost 700,000 tweets in the corpus, the two language detectors agreed on 69.4% of the tweets as being Frisian (21.1%), Dutch (20.3%), or undetermined (28%). One-third of the total was classified as Frisian by at least one of the detectors. This shows that combining the two language detectors resulted in a high number of Frisian tweets that would otherwise have been discarded. The language detectors disagreed on almost a third of the tweets (30.6%). This might be explained by the close relation between the two languages, and the many code switches and phonetic spelling that were used by the Twitterers.

As mentioned in Section Data cleaning and Coding we found 5,559 tokens in the manual coding. Most tokens (5,314) came from tweets that were detected as Frisian by one or both language detectors. Additionally, the category that was detected as "undetermined" contained a significant number of tokens, i.e., 238 as well. In contrast, the 13,932 tweets that were classified as Dutch by both language detectors had seven tokens from seven tweets (0.05%). In other words, in future analyses it is beneficial to combine the two language detectors and perform a manual check on the Frisian and undetermined tweets and discard the Dutch ones. One would then only miss a small number of tokens and gain a lot of time.

The Final Set of Tweets With at Least One of the Target Variables (5,395 Tweets, 5,559 Tokens, 159 Twitterers)

The analysis of linguistic variables is often hampered by the unequal distribution of the variable over linguistic contexts or speakers (the frequency problem), the entanglement of linguistic factors resulting in (in)frequent combinations of these factors (the co-occurrence problem) and the existence of (groups of) speakers showing linguistically different patterns of variation (the interaction problem) (Van de Velde and van Hout, 2000).

Table 4 presents the frequency distribution of the variants of the four target variables (*dy't*), (*dêr't*), (*wêr't*) and (*wa't*). The tokens are unequally distributed over the target variables. More than two-thirds of the tokens, i.e., 3,807 (68.4%), are variants of (*dy't*), and the remaining part consists of variants of the other three target variables (co-occurence problem). When looking at the (*dy't*)-variable in more detail we see several variants of the t-full forms, i.e., *dy't*, **die't*, *dy'tst*, and **die'tst*. In total we see

TABLE 4 | Distribution of Frisian relative pronouns (*dy't*), (*dêr't*), (*wêr't*), and (*wa't*) in the final data set ($n = 5{,}559$, distributed over 159 Twitterers).

Target variable	t-full forms	*n*	t-less forms	*n*
(dy't)	*dy't*	3278	**dy*	136
	**die't*	55	**die*	286
	dy'tst	4	**dy(')st*	24
	**die'tst*	0	**die(')st*	24
(dêr't)	*dêr't*	408	**dêr*	0
	**der't*	23	**der*	0
	dêr'tst	2	**dêr(')st*	15
	**der'tst*	0	**der(')st*	1
(wêr't)	*wêr't*	521	**wêr*	78
	**wer't*	52	**wer*	98
	wêr'tst	1	**wêr(')st*	21
	**wer'tst*	0	**wer(')st*	30
(wa't)	*wa't*	334	**wa*	139
	**wie't*	1	**wie*	19
	wa'tst	3	**wa(')st*	4
	**wie'tst*	0	**wie(')st*	2
Total		**4682**		**877**

**Ungrammatical variant.*

TABLE 5 | Characteristics of the Twitterers: gender, birth year (per decade) and spelling style (standard or phonetic Frisian) per number of tokens ($n = 5{,}559$) and Twitterers ($N = 159$).

Variable	Categories	*n* tokens	*N* twitterers
Gender	Male	2,779	81
	Female	2,780	78
Birth year	1941–1950	870	9
	1951–1960	1,431	22
	1961–1970	1,484	24
	1971–1980	890	21
	1981–1990	388	29
	1991–2000	496	54
Spelling style	Standard Frisian	4,863	83
	Phonetic Frisian	696	76

3,337 realizations of these t-full forms of *dy't*), which comes down to 87.7%, whereas 470 realizations (12.3%) are t-less variants of (*dy't*). Although most realizations of (*dy't*) are t-full, the variant where the Dutch spelling (**die*) is used and the variants that have inflection of second person singular (**dy(')st* and **die(')st*) mostly have a t-less form.

As for the second most frequent target variable (*wêr't*) ($n = 801$; 14.4%), we see in **Table 4** that most realizations are t-full forms, but it seems that the variants without accent, **wer*, and with the suffix of second person singular *–st*, are more frequently realized as t-less forms. The variable (*wa't*) ($n = 502$; 9.0%) shows a similar pattern as (*dy't*): most realizations are t-full forms, but the variants where the Dutch spelling (**wie*) is used and the variants inflection of second person singular (**wa(')st* and **wie(')st*) mostly have a t-less form. The variable (*dêr't*) has the lowest frequency ($n = 449$; 8.1%). Note also that (*dêr't*) is always used with t-full forms, except when it is inflected with the second person singular suffix. In that case, the Twitterers mostly use a t-less form (see **Table 4**).

On average, there are 35 occurrences of the variable per Twitterer. **Table 5** presents the distribution of the number of tokens, and Twitterers, split up for gender, birth year (per decade), writing style (phonetic or standard Frisian). There are no differences between men and women. The tokens show up most frequently in the tweets of Twitterers born between 1951 and 1970. In tweets of Twitterers born after 1980 the frequency of the number of tokens is much lower. In contrast, the younger Twitterers are overrepresented in this data set compared to the older Twitterers. This is in line with findings from Blank (2016) that Twitterers are generally younger than the offline population. The unequal distribution of the variable over the birth cohorts might hamper a robust analysis of this data set from the perspective of language change. 87.5% of the tokens showed up in tweets in standard Frisian orthography, only 12.5% in tweets with phonetic spelling. This supports previous observations that in CMC Twitter mainly shares characteristics with traditional writing styles (Verheijen, 2018).

Figure 5 shows the number of Twitterers distributed over the percentage of using t-full forms in their tweets. Forty-nine Twitterers (30.8%) never use the t-full forms and 36 Twitterers (22.6%) always use the t-full forms. Seventy-four Twitterers (46.5%) vary in their use of t-full and t-less forms. **Figure 6** illustrates in more detail the unequal frequency distribution of the tokens. It should also be noted that 21 Twitterers use the target variable only once in their tweets posted in the decade 2010–2019. Such an unequal distribution of tokens (frequency problem) might hamper a solid analysis of the factors influencing the use of t-full variants. It is likely that the Twitterers with many tokens are the most-skilled writers, with solid knowledge of Frisian normative rules and standard orthography (interaction problem).

When ranking the final set of tweets with at least one of the target variables by birth year (see **Table A** that is uploaded as **Supplementary Material**), we see that without exception the Twitterers born after 1988 stop tweeting after 2014. The fact that the older Twitterers produce most of the tokens and the younger Twitterers leave Twitter after 2014, which is a problem for the analysis of the interaction of time and age in our panel study.

In an attempt to cope with the above problems related to the distribution of the tokens, the Twitterers were split in three groups. A small group of Twitterers ($n = 14$), FreqvarH, realizes two-thirds of the total amount of tokens in the data set (see **Table 6** and **Figure 6**). Most of the Twitterers ($n = 118$) produce a relatively low number of tokens of the variable (FreqvarL, range 1–21). We also created an intermediate group, FreqvarM ($n = 27$, range = 23–88). In the results section we will present separate analyses for these groups.

Analysis of Tweets of Final Data Set

The data were analyzed using a mixed-effects logistic regression model in R (The R Foundation for Statistical Computing,

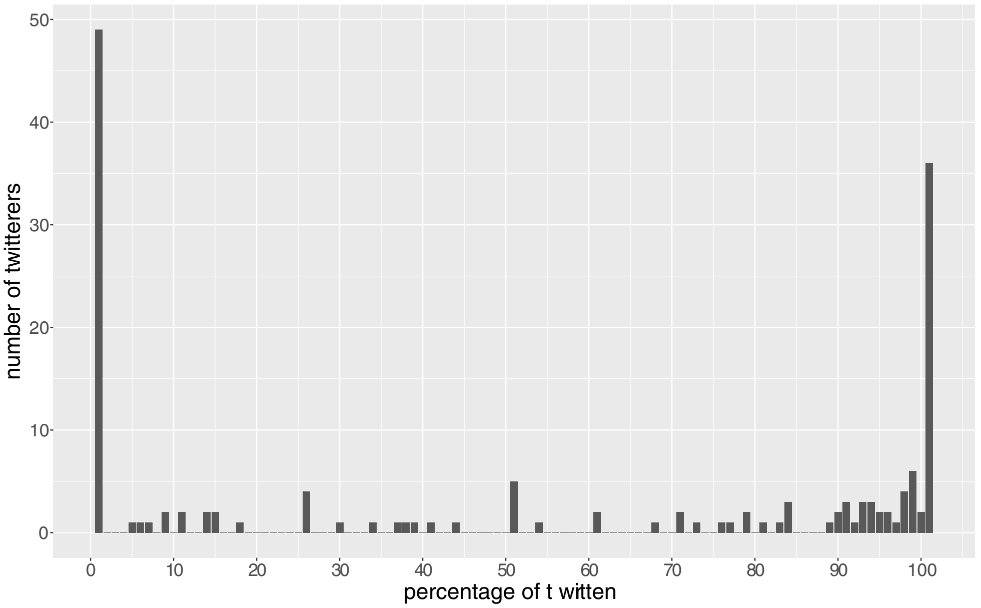


FIGURE 5 | Distribution of number of Twitterers per percentage of t-full forms used (N = 159).

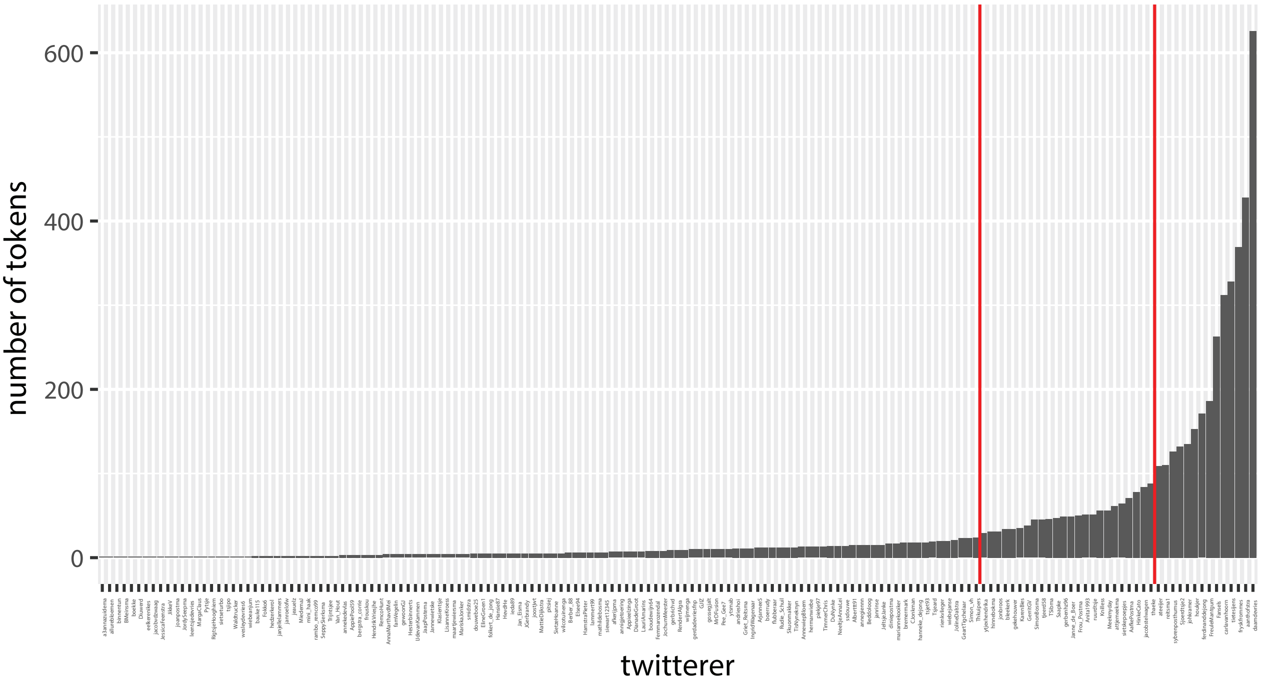


FIGURE 6 | Number of tokens per Twitterer (N = 159).

http://CRAN.R-project.org/) by applying the glmer function in the lme4 package (Bates et al., 2015). This function uses a combination of Nelder-Mead and bobyqa as optimizer. The model did not converge when it included the variable "pronoun." Therefore, we used the optimizer nlminb from the R package optimx, which solved the conversion issues. The code of nlminb() was written by David Gay at Bell Labs and part of the Fortran library (Fox et al., 1978).

The dependent variable was (t) with 0 indicating a t-less form and 1 indicating a t-full form. The analysis started with an initial

TABLE 6 | Distribution of number of Twitterers ($N = 159$) and tokens ($n = 5{,}559$), split up in three frequency groups.

	Range per Twitterer	*n* tokens	*N* Twitterers
FreqvarL	1–21	818	118
FreqvarM	23–88	1,293	27
FreqvarH	106–626	3,448	14

TABLE 7 | Optimal model for substitution of Frisian relative pronouns *dy't*, *dêr't*, *wêr't,* and *wa't* in Twitter data for FreqvarL, Twitterers who infrequently produced the variable (range = 1–21) in their tweets ($n = 818$, $N = 118$).

	Estimate	S.E.	*Z* value	Pr (>|z|)
Intercept	3.65	1.18	3.08	$p < 0.001$
Birth year	−1.60	0.40	−3.96	$p < 0.00$
#words	0.31	0.15	2.06	$p < 0.05$
Phonetic spelling	−2.98	0.80	−3.73	$p < 0.001$
(dy't)	−0.36	1.13	−0.32	n.s.
(wa't)	−1.38	1.17	−1.18	n.s.
(wêr't)	−1.29	1.16	−1.11	n.s.

model that included the following fixed factors: birth year and gender of the Twitterers, the year in which the tweet was posted, the spelling style, the pronoun, the number of words used in the tweet, the number of tokens per Twitterer, the percentage of Frisian used in the tweets, and the presence of the suffix *–st*. The percentage of Frisian per Twitterer is calculated as the number of Frisian tweets divided by the total number of tweets multiplied by 100. The number of (Frisian) tweets is calculated on the basis of the full data set, i.e., the set which contains all tweets regardless of the presence of any of the target words. A tweet is considered Frisian when it is detected as Frisian by any of the two language detectors that we used (see Section Automatic Detection of Frisian Tweets).

The Twitterers were included as a random factor, to control for individual differences. The variable pronoun was included as random slope. Starting with the initial model, backward analysis was acquired to obtain the best model that had a significant improvement of the Akaike Information Criterion (AIC). We followed this procedure for each of the three groups of Twitterers (low, moderate and high frequency of the variable) from **Table 6**. The optimal models from these analyses are presented in the next section.

RESULTS

We present separate analyses for the groups with low, moderate, and high frequency of the variable.

The optimal model for the FreqvarL-Twitterers, i.e., Twitterers who use a small number of tokens (1 up to 21), is presented in **Table 7**. Younger Twitterers substitute the Frisian relative pronouns significantly more with the t-less forms of relative pronouns, compared to older Twitterers. Further, longer tweets (in terms of number of words), have significantly more t-less forms. Also, phonetic spelling is an important factor. Twitterers who tweet in phonetic Frisian use significantly more t-less forms. Pairwise comparisons for pronouns show that significantly more t-less forms are found for *dy't* compared to *wêr't* ($z = 2.69$, $p < 0.05$).

TABLE 8 | Optimal model for substitution of Frisian relative pronouns *dy't*, *dêr't*, *wêr't,* and *wa't* in Twitter data for FreqvarM, Twitterers who moderately produced the variable (range = 23–88) in their tweets) ($n = 1{,}293$, $N = 27$).

	Estimate	S.E.	*Z* value	Pr (>|z|)
Intercept	6.74	2.68	2.52	$p < 0.05$
Year of posting	0.37	0.17	2.07	$p < 0.05$
Phonetic spelling	−6.15	0.86	−7.14	$p < 0.001$
Suffix *-st*	−16.94	5.00	−3.39	$p < 0.001$
(dy't)	−2.30	2.65	−0.87	n.s.
(wa't)	−4.46	2.67	−1.67	n.s.
(wêr't)	−5.02	2.66	−1.89	n.s.

Table 8 presents the optimal model for FreqvarM-Twitterers who moderately use the variable (23 up to 88) in their tweets. Within this model the suffix *–st* is also included as a random slope. The model shows that within this group the year of posting is a significant factor. The more recent the posting, the more t-full forms were used. Additionally, the Twitterers using phonetic spelling used significantly more t-less forms compared to those using standard Frisian spelling. The addition of the suffix *–st*, when the subject in the relative clause is second person singular, also triggers the use of t-less relative pronouns. Pairwise comparisons for pronouns show that significantly more t-less forms are found for *dy't* and *wa't* ($z = 5.31$, $p < 0.001$) and *wêr't* ($z = 7.48$, $p < 0.001$). Variables such as birth year, gender, the number of words in a tweet or the percentage Frisian used in the tweets were not included in this model.

Table 9 presents the optimal model for the FreqvarH-Twitterers, i.e., Twitterers with a high frequency of the variable (range: 106–626) in their tweets. A fixed effect is found for the variable count. This variable represents the number of tweets that a Twitterer posted between January 1, 2010 and December 31, 2019. It means that Twitterers who posted more tweets used more t-full forms. Additionally, when the suffix *–st* is cliticized to the relative pronoun, the pronoun itself shows up more frequently as a t-less form. The table further shows an effect for pronoun. Pairwise comparisons show that significantly more t-less forms are found for *dy't* and *wa't* ($z = 6.85$, $p < 0.001$) and *wêr't* ($z = 4.50$, $p < 0.01$), and more t-full forms for *dêr't* compared to *wa't* ($z = 3.70$, $p < 0.01$) and *wêr't* ($z = 2.93$, $p < 0.05$). The variables birth year, gender, year of posting, number of words of a tweet, spelling style, or the percentage of Frisian in the tweet do not significantly contribute to the optimal model.

DISCUSSION AND CONCLUSION

The current study investigated a change in progress in Frisian based on Twitter data: the substitution of t-full relative pronouns *dy't*, *dêr't*, *wêr't,* and *wa't* with their t-less counterparts. The aim of the study was threefold. First, we wanted to explore the issues in gathering a Twitter corpus of a low-resource language such as

TABLE 9 | Optimal model for substitution of Frisian relative pronouns *dy't*, *dêr't*, *wêr't,* and *wa't* in Twitter data for FreqvarH, i.e., Twitterers who frequently produced the variable (range = 106–626) in their tweets) (n = 3,448, N = 14).

	Estimate	S.E.	*Z* value	Pr (>|z|)
Intercept	6.51	1.08	6.01	$p < 0.001$
Count	1.21	0.50	2.42	$p < 0.05$
Suffix *-st*	−8.39	0.94	2.06	$p < 0.05$
(dy't)	−1.78	1.01	−1.76	n.s.
(wa't)	−3.81	1.03	−3.70	$p < 0.001$
(wêr't)	−3.00	1.02	−2.93	$p < 0.01$

Frisian. Second, we wanted to get more insight in the validity of Twitter data for the study of language change in progress. Third, we tried to enhance our insight in a vigorous change in progress.

Collecting Data

The collection of Frisian Twitter data turned out to be a complex process. Multiple requests to get permission to retrieve tweets from Twitter for linguistic research did not result in an answer from Twitter. Consequently, we used the GetOldTweets3-script to retrieve the tweets of a fixed set of individual Twitter accounts of Frisian Twitterers that were previously identified and selected in another project (Jongbloed-Faber et al., 2017). A corpus of almost 700,000 predominantly Frisian and Dutch tweets posted in the decade 2010–2019 was collected. After automatically selecting the possible realizations of the variables and a check analysis, we ended up with the final data set of 5,395 Frisian tweets with one or more realizations of the variables, which is a fraction (0.8%) of the tweets from the entire Twitter data set. Although being recognized as the second official language in the province, Frisian is not omnipresent in the written domain and only a small proportion of the 450,000 speakers of Frisian write Frisian. Hence, data sets and corpora of such languages are much smaller than the ones of majority or medium-sized languages. Furthermore, many Twitterers of a minority language are bilingual and tweet in the majority language (as well). Evidence from a previous study (Jongbloed-Faber et al., 2017) showed that the Twitterers from our data set used (some) Frisian in their tweets, next to Dutch (or occasionally another language). Jongbloed-Faber et al. (2016) also pointed out that 65% of the Frisian teenagers never use Frisian in their tweets. So, the Twitterers of our data set might give a slightly distorted view of the language use of the average Frisian Twitterers who predominantly use Dutch.

Detection of the Language

The automatic identification of tweets as Frisian or Dutch was not very successful. Almost one-third of the tweets was classified as undetermined and these tweets contained a significant number of tokens of our linguistic variable. Frisian and Dutch are closely related languages, and the Frisian lexicon contains a lot of Dutch loans. The fact that a large part of the Frisian tweets is written in phonetical spelling, heavily influenced by Dutch spelling conventions, make automatic distinction between Frisian and Dutch tweets even more difficult. Furthermore, the corpora behind the language detector textcat are relatively small, which makes such a detector less performant. Like most minority and smaller languages, Frisian is technologically a low-resource language and at the time of our research a POS-tagger for Frisian was not available. A POS-tagger for Frisian would have made it easier to distinguish between *dy* used as t-less relative pronoun or demonstrative pronoun, *wêr* used as t-less relative pronoun or interrogative pronoun, *wer* as relative pronoun or adverb, *wa* as relative pronoun or interrogative pronoun, *wie* as (Dutch) relative pronoun or (Frisian) inflection of *wêze* 'to be'. Consequently, the 100,365 tweets had to be analyzed manually to distinguish the target variables from other words.

Analysis and Imbalance of Data

Three of the four relative pronouns in our study had a low frequency in comparison with the fourth one. This co-occurrence problem makes it difficult to study the role of this linguistic factor and might explain why it does not show up as a significant factor in the low and medium frequency groups.

Differences in the quantity of tweets is not a problem, if a comparable sample of tokens per individual can be selected. However, two-thirds of the tokens of our variable are produced by less than one-tenth of the Twitterers in our corpus, and most of the Twitterers produce a low number of tokens in the decade we were able to track their tweets. This unequal distribution of tokens is problematic for a panel study of language change in progress.

The final data set appeared to be even more biased. In our panel study, we observe a strong decrease in the use of the medium. Most tweets are posted between 2011 and 2013. After 2014 there is a rapid decline in the number of tweets. Striking is that all Twitterers in our data set born after 1988, without any exception, stopped posting on Twitter after 2014. This is a problem for panel studies like this one, especially in low resource languages where the amount of data is rather limited. Furthermore, most tokens of our variable are produced by Twitterers from older generations, hampering an analysis of the data set in apparent time, and the interaction of age and period.

Findings

The shortcomings of the data set did not imply that we could not refine the existing insights in this change in progress, since the data set showed two interesting observations. The first observation concerned the target variable *dêr't*. A previous study on radio speech data showed that the target variable *dêr't* was mostly found in scripted radio speech and almost always in t-full form (Dijkstra et al., 2019). The current study demonstrated that in tweets, the t-full form is always used in *dêr't* (unless this relative pronoun is inflected with second person singular suffix –*st*). This suggests that the relative pronoun *dêr't* is part of written rather than of oral Frisian. A second observation concerns the suffix –*st*. The suffix –*st* seems to trigger the t-less form of all target variables predominantly in tweets from the two most active groups of Twitterers. This might be explained by the observation that the /t/ before the –*st* suffix is usually not pronounced (Hoekstra, 1985). Due to the bias in the data set, we have to be careful in generalizing our findings. They need to be confirmed on the basis of additional analysis of spoken and written corpora.

When studying language change in panel studies one needs to monitor individuals over a period of time. The instability in token production by individuals and the general decline of the medium, especially amongst young Twitterers, make it hard to demonstrate language change in progress in real time. Our analyses were further hampered by the fact that the variables had a low frequency and were unequally distributed over Twitterers. The fact that the language under investigation was a low-resource language, made the search and analysis even more challenging. In conclusion, for low-frequency variables in low-resource languages, Twitter is unlikely to be an appropriate source for quantitative sociolinguistic studies of language change in progress.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable data included in this article as the data were anonymized.

AUTHOR CONTRIBUTIONS

JD, WH, LJ-F, and HV all collaborated on the conception of the study. The design was developed by JD, WH, and HV. Data was collected by LJ-F and WH and prepared by WH. Data was manually analyzed by JD. Statistical analyses were performed by WH. A first draft of the paper was written by JD and HV. WH and LJ-F (co-)wrote sections of the paper. JD and HV prepared the final version of the manuscript which was read, revised, and approved by WH and LJ-F. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We are thankful for using the set of individual Twitter accounts derived from the data set collected within the Twidentity project (supported by NWO, KIEM 314-98-008), a joint project of University of Twente, Meertens Institute and Fryske Akademy.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.644554/full#supplementary-material

REFERENCES

Androutsopoulos, J. (2006). Introduction: sociolinguistics and computer-mediated communication. *J. Sociolinguist.* 10, 419–438. doi: 10.1111/j.1467-9841.2006.00286.x

Androutsopoulos, J. (2014). Moments of sharing: entextualization and linguistic repertoires in social networking. *J. Pragmat.* 73, 4–18. doi: 10.1016/j.pragma.2014.07.013

Baron, N. S. (2010). Discourse structures in Instant Messaging: the case of utterance breaks. *Language@Internet* 7article 4.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Blank, G. (2016). The digital divide among Twitter users and its implications for social research. *Soc. Sci. Comput. Rev.* 35, 679–697. doi: 10.1177/0894439316671698

Bleaman, I. L. (2020). Implicit standardization in a minority language community: real-time syntactic change among Hasidic Yiddish Writers. *Front. Artif. Intell.* 3:35. doi: 10.3389/frai.2020.00035

Brouwer, J. H. (1959). Mei té of sûnder té [with /t/ or without /t/]. *De Pompebledden* 30, 62–64.

Cavnar, W. B., and Trenkle, J. (1994). *N-Gram-Based Text Categorization*. Ann Arbor MI: Environmental Research Institute of Michigan

Chignell, H. (2009). *Key Concepts in Radio Studies*. London: Sage.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139164771

Cunliffe, D., Morris, D., and Prys, C. (2013). Young bilinguals' language behaviour in Social Networking Sites: the use of Welsh on Facebook. *J. Comput. Med. Commun.* 18, 339–361. doi: 10.1111/jcc4.12010

Cutler, C., and Røyneland, U. (eds.). (2018). *Multilingual Youth Practices in Computer Mediated Communication*. Cambridge: Cambridge University Press. doi: 10.1017/9781316135570

De Decker, B. (2014). *"De chattaal van Vlaamse tieners: Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur*. [The chat language of Flamish teenagers: A language geographic analysis of Flamish (sub)standardising processes against the background of the international chat culture]." (Ph.D. dissertation). University of Antwerp, Antwerpen, Belgium.

De Decker, B., VandeKerckhove, R., and Sandra, D. (2016). when two basic principles class: about the validity of written chat language as a research tool for spoken language variation. Flemish Chatspeak as a Test Case. *J. Lang. Contact* 9, 101–129. doi: 10.1163/19552629-00901005

De Haan, G. J. (2001). *Grammar of Modern West Frisian*. University of Groningen, Groningen, the Netherlands.

Dijkstra, J., Heeringa, W., Yilmaz, E., van den Heuvel, H., van Leeuwen, D., and Van de Velde, H. (2017). "A real time study of contact-induced language change in Frisian relative pronouns." in *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017* ed E. Babatsouli (Chania: Institute of Monolingual and Bilingual Speech), 113–119.

Dijkstra, J., Heeringa, W., Yilmaz, E., van den Heuvel, H., van Leeuwen, D., Van de Velde, H., et al. (2018). "Tracking real time language change in relative pronouns in spoken West-Frisian," in *Cross-Linguistic Research in Monolingual and Bilingual Speech,* ed E. Babatsouli (Chania: ISMBS), 93–109.

Dijkstra, J., Heeringa, W., Yilmaz, E., van den Heuvel, H., van Leeuwen, D., and Van de Velde, H. (2019). "Language change caught in the act: a case study of t-deletion in Frisian relative pronouns." in *Language Variation. European Perspectives VII. Selected Papers from the Ninth International Conference on Language Variation in Europe (ICLaVE9),* eds J. A. Villena-Ponsoda, F. Díaz Montesinos, A. M. Ávila-Muñoz and M. Vida-Castro (Amsterdam: John Benjamins), 81–101. doi: 10.1075/silv.22.05dij

Eisenstein, J. (2013). "Phonological factors in social media writing," *Proceedings of the Workshop on Language in Social Media (LASM 2013)* (Portland, OR; Association for Computational Linguistics), 11–19.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *J. Sociolinguist.* 19, 161–188. doi: 10.1111/josl.12119

Fox, P. A., Hall, A. D., and Schryer, N. L. (1978). The PORT mathematical subroutine library. *ACM Trans. Math. Softw.* 4:104E126. doi: 10.1145/355780.355783

Glicoric, C., Anderson, A., and West, R. (2020). Adoption of Twitter's new length limit: Is 280 the New 140? Available online at: https://arxiv.org/abs/2009.07661 (accessed February 17, 2021).

Gorter, D. (2001). "Extend and position of West Frisian." in *Handbuch des Friesischen/Handbook of Frisian Studies,* eds H. H. Munske, N. Arhammer, V. F. Faltings, J. F. Hoekstra, O. Vries, A. G. H. Walker et al. (Tübingen: Max Niemeyer),73–83.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., Guo, D. (2019). Mapping lexical dialect variation in Brittish English Using Twitter. *Front. Artif. Intell.* 2:11. doi: 10.3389/frai.2019.00011

Grondelaers, S. A., van Hout, R. W. N. M., and van Halteren, H. (2017). "Hun twitteren. Tweets als bron voor onderzoek naar syntactische taalvariatie [They tweet. Tweets as source for research into syntactic language variation]," in *Taalvariatie en sociale media,* eds V. De Tier, T. Wijngaard and A. Ghyselen (Leiden: Stichting Nederlandse Dialecten), 65–72.

Herring, S. C. (ed.). (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives. Pragmatics and Beyond Series*. Amsterdam: John Benjamins. doi: 10.1075/pbns.39

Herring, S. C. (2001). "Computer-mediated discourse." in *The Handbook of Discourse Ana.lysis,* eds D. Schiffrin, D. Tannen and H. Hamilton (Oxford: Blackwell Publishers), 612–634.

Herring, S. C., and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *J. Sociolinguist.* 10, 439–459. doi: 10.1111/j.1467-9841.2006.00287.x

Hoekstra, J. (1985). T-deletion before suffix-initial st in Modern West Frisian. *Nowele* 5, 63–76. doi: 10.1075/nowele.5.04hoe

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., and Feinerer, I. (2013). The textcat package for nGgram based text categorization in R. *J. Stat. Softw.* 52, 1–17. doi: 10.18637/jss.v052.i06

Jongbloed-Faber, L. (2015). *Friezen op sosjale media: Rapportaazje ûndersyk Taalfitaliteit II [Frisians on social media: Report Research Language Vitality II].* Ljouwert, the Netherlands: Fryske Akademy—Mercator European Research Centre on Multilingualism and Language Learning.

Jongbloed-Faber, L., Van de Velde, H., van der Meer, C., and Klinkenberg, E. L. (2016). Language use of Frisian bilingual teenagers on social media. *Treballs de Sociolingüística Catalana* 26, 27–54. doi: 10.2436/20.2504.01.107

Jongbloed-Faber, L., van Loo, J., and Cornips, L. (2017). Regional languages on Twitter. A comparative study between Frisian and Limburgish. *Dutch J. Appl. Linguist.* 6, 174–196. doi: 10.1075/dujal.16017.jon

Jonkman, R., and Versloot, A. (2018). *The Story of Frisian in Multilingual Friesland*. Ljouwert/Leeuwarden: Afûk.

Klinkenberg, E., Jonkman, R., and Stefan, N. (2018). *Taal yn Fryslân. De folgjende generaasje* [Language in Fryslân. The Next Generation]. Ljouwert: Fryske Akademy.

Munske, H. H., Århammar, N., Faltings, V., Hoekstra, J., Vries, O., Walker, A., et al. (eds.). (2001). *Handbook of Frisian Studies*. Tübingen: Max Niemeyer Verlag. doi: 10.1515/9783110946925

Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think I am? A study of language and age in Twitter," In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (Cambridge, MA), 439–448.

Ooms, J. (2020). *cld3: Google's Compact Language Detector 3. R package version 1.3.* Available online at: https://CRAN.R-project.org/package=cld3 (accessed February 5, 2021).

Plester, B., Wood, C., and Bell, V. (2008). Txt Msg n school literacy: does texting and knowledge of text abbreviations adversely affect children's literacy attainment? *Literacy* 42, 137–144. doi: 10.1111/j.1741-4369.2008.00489.x

Popkema, J. (2018). *Grammatica Fries.* Leeuwarden: Afûk.

Reershemius, G. (2017). Autochthonous heritage languages and social media: writing and bilingual practices in Low German on Facebook. *J. Multilingual Multicult. Dev.* 38, 35–49. doi: 10.1080/01434632.2016.1151434

Stæhr, A. (2015). Reflexivity in Facebook interaction: enregisterment across written and spoken language practices. *Discourse Context Media* 8, 30–45. doi: 10.1016/j.dcm.2015.05.004

Taalportaal|Relative pronouns (2018). Available online at: https://taalportaal.org/taalportaal/topic/pid/topic-13998813311730000 (accessed February 17, 2021).

Thurlow, C., Lengel, L., and Tomic, A. (2004). *Computer Mediated Communication: Social Interaction and the Internet*. Thousand Oaks, CA: Sage.

Van Blom, P. H. (1889). *Beknopte Friesche Spraakkunst voor den tegenwoordige tijd.* Joure: R.P. Zijlstra.

Van de Velde, H., and van Hout, R. (2000). "N-deletion in reading style," in *Linguistics in the Netherlands 2000,* eds H. de Hoop and T. van der Wouden (Amsterdam: John Benjamins), 209–219. doi: 10.1075/avt.17.20van

Van der Meer, G. (1991). The subclause signal't in Frisian. Its origin and function. *Leuvense Bijdragen* 80, 43–59.

Van der Woude, G. (1960). "Oer it gebrûk fan't by bynwurden [On the use of't in conjunctions]," in *Fryske Stúdzjes: Fryske Stúdzjes oanbean oan Prof. Dr. J.H. Brouwer op syn sechstichste jierdei 23 augustus 1960,* eds K. Dykstra, K. Heeroma, W. Kok and H. T. J. Miedema (Assen: Van Gorcum), 335–343.

Vandekerckhove, R. (2006). Chattaal, tienertaal en taalverandering: (sub)standaardiserings-processen in Vlaanderen [Chat language, teenagers language and language change: (sub)standardizing processes in Flanders]. *Handelingen der Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis* 59, 139–158.

Vandekerckhove, R., and Nobels, J. (2010). Code eclecticism: linguistic variation and code alternation in the chat language of Flemish teenagers. *J. Sociolinguist.* 14, 657–677. doi: 10.1111/j.1467-9841.2010.00458.x

Verheijen, L. (2017). "WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication." in *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World,* eds D. Fišer and M. Beißwenger (Ljubljana: Ljubljana University Press), 72–101.

Verheijen, L. (2018). *Is Textese a Threat to Traditional Literacy? Dutch Youths' Language Use in Written Computer-Mediated Communication and Relations With Their School Writing* (Ph.D. dissertation), University of Nijmegen, the Netherlands

Wagner, S. E., and Buchstaller, I. (eds.). (2017). *Panel Studies of Variation and Change*. New York, NY: Routledge. doi: 10.4324/9781315696591

Wang, Z., Hale, S. A., Adelani, D., Grabowicz, P. A., Hartmann, T., Flöck, F., et al. (2019). "Demographic interference and representative population estimates from multilingual social media data," in *Proceedings of WWW'19 The World Wide Web Conference* (San Francisco, CA: ACM), 2056–2067. doi: 10.1145/3308558.3313684

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/frai.2021.648543

Considering Performance in the Automated and Manual Coding of Sociolinguistic Variables: Lessons From Variable (ING)

Tyler Kendall[1]*, *Charlotte Vaughn*[1,2], *Charlie Farrington*[1,3], *Kaylynn Gunter*[1], *Jaidan McLean*[1], *Chloe Tacata*[1] and *Shelby Arnson*[1]

[1] Linguistics Department, University of Oregon, Eugene, OR, United States, [2] Language Science Center, University of Maryland, College Park, MD, United States, [3] English Department, North Carolina State University, Raleigh, NC, United States

Impressionistic coding of sociolinguistic variables like English (ING), the alternation between pronunciations like *talkin'* and *talking*, has been a central part of the analytic workflow in studies of language variation and change for over a half-century. Techniques for automating the measurement and coding for a wide range of sociolinguistic data have been on the rise over recent decades but procedures for coding some features, especially those without clearly defined acoustic correlates like (ING), have lagged behind others, such as vowels and sibilants. This paper explores computational methods for automatically coding variable (ING) in speech recordings, examining the use of automatic speech recognition procedures related to forced alignment (using the Montreal Forced Aligner) as well as supervised machine learning algorithms (linear and radial support vector machines, and random forests). Considering the automated coding of pronunciation variables like (ING) raises broader questions for sociolinguistic methods, such as how much different human analysts agree in their impressionistic codes for such variables and what data might act as the "gold standard" for training and testing of automated procedures. This paper explores several of these considerations in automated, and manual, coding of sociolinguistic variables and provides baseline performance data for automated and manual coding methods. We consider multiple ways of assessing algorithms' performance, including agreement with human coders, as well as the impact on the outcome of an analysis of (ING) that includes linguistic and social factors. Our results show promise for automated coding methods but also highlight that variability in results should be expected even with careful human coded data. All data for our study come from the public Corpus of Regional African American Language and code and derivative datasets (including our hand-coded data) are available with the paper.

Keywords: English variable (ING), impressionistic coding, automated coding, classification, machine learning, forced alignment, sociolinguistic variables

OPEN ACCESS

Edited by:
Jussi Karlgren,
Spotify, Sweden

Reviewed by:
Helmer Strik,
Radboud University
Nijmegen, Netherlands
Chantal Gratton,
Stanford University, United States

****Correspondence:***
Tyler Kendall
tsk@uoregon.edu

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *31 December 2020*
Accepted: *23 March 2021*
Published: *29 April 2021*

Citation:
Kendall T, Vaughn C, Farrington C, Gunter K, McLean J, Tacata C and Arnson S (2021) Considering Performance in the Automated and Manual Coding of Sociolinguistic Variables: Lessons From Variable (ING). Front. Artif. Intell. 4:648543. doi: 10.3389/frai.2021.648543

INTRODUCTION

Since the earliest days of variationist sociolinguistic research (e.g., Labov, 1963, 1966; Wolfram, 1969; Trudgill, 1974), variable pronunciations in speech collected from communities of speakers have been the basis for much research into the principles and processes of language variation and change. A key methodology in this tradition involves the impressionistic coding of sociolinguistic variables (Wolfram, 1993) – such as determining whether a post-vocalic /r/ was vocalized or rhotic (e.g., *guard* as [ga:d] vs. [gard]) or a word final *-ing* in words like *talking* was produced as *-in* or *-ing* – and making quantitative comparisons within and across speakers in the use of these variables. This work has led to key observations about the *orderly heterogeneity* of language (Weinreich et al., 1968), the systematicity underlying the social and linguistic bases for language variation and change. One bottleneck in sociolinguistic research, especially as opportunities increase to study larger and larger collections of spoken language, has been the immense work that goes into coding sociolinguistic variables. While research on some variable phenomena, like vowels and sibilants (e.g., Labov et al., 1972; Stuart-Smith, 2007; see Kendall and Fridland, 2021), has been advanced by acoustic phonetic analysis, many pronunciation features of interest to sociolinguists, like coronal stop deletion (e.g., Guy, 1980; Hazen, 2011), variable rhoticity or *r*-lessness (e.g., Labov, 1966), final stop devoicing (Farrington, 2018, 2019), and velar nasal fronting or variable (ING) (e.g., Tagliamonte, 2004; Hazen, 2008), the variable under focus in this paper, have continued to rely on careful impressionistic coding by analysts.

Techniques for automating the measurement and coding of sociolinguistic data have been on the rise for the past couple of decades, and parallel developments in automation for other areas of the phonetic sciences (such as the phonetic transcription of large corpora; see e.g., Van Bael et al., 2007). Sociophonetic analyses of vowels, in particular, have seen major methodological advances via popular software like the Forced Alignment and Vowel Extraction suite (FAVE; Rosenfelder et al., 2014, see e.g. Labov et al., 2013 for a large-scale example of its use), and efforts have been ongoing to automate other sociophonetic workflows (Sonderegger et al., in progress). These methods have almost entirely replaced the impressionistic coding for such features, which was a mainstay of early sociolinguistic research (e.g., Labov, 1963, 1966; Trudgill, 1974). The success of these methods for particular features, and the degree of appropriateness of acoustic analysis (as opposed to impressionistic coding) more generally, has hinged on the field's ability to identify acoustic dimensions that relate reliably to the auditory impressions of listeners. Features like vowels and sibilants have relatively straightforward acoustic cues, and for features like these, the field has moved over time to view acoustic measures as more useful than the impressionistic coding of analysts (though we further discuss the implications of such a move, which removes the consideration of auditory importance, in the section Determining the Realization of Pronunciation Variables). This paper focuses on the case of sociolinguistic variables that do not have straightforward acoustic cues, like variable (ING), which have remained the domain of impressionistic, categorical coding.

Corpus phonetic approaches (Liberman, 2019; Kendall and Fridland, 2021: chapter 8) have been growing in popularity and a turn to "bigger data" somewhat necessitates an ability to code data in more cost- and time-efficient ways. Thus, the application of automated approaches to the coding of sociolinguistic variables that typically require manual categorical coding represents an important area for methodological improvement. Some work has engaged in this problem, especially in the phonetic sciences broadly (e.g., Van Bael et al., 2007; Schuppler et al., 2011), although relative to the automatic measurement of features like vowels, as just discussed, efforts for the coding of many sociolinguistic variables are not as advanced. To our knowledge, studies thus far have explored automated techniques for coding the deletion of /n/, /r/, and /t/, as well as schwa deletion and insertion, in Dutch (Kessens et al., 1998; Wester et al., 2001) and, for English, /l/ darkness (Yuan and Liberman, 2009, 2011a), post-vocalic *r*-lessness (McLarty et al., 2019; Villarreal et al., 2020), features of /t,d/ (Bailey, 2016; Villarreal et al., 2020), and, the focus of this paper, variable (ING) (Yuan and Liberman, 2011b), but such work is in its relative infancy and these prior studies, as well as the current paper, set the stage for further advancement.

Two broad approaches have been proposed for the automatic coding of categorical pronunciation features. The first, as proposed and implemented by Kessens et al. (1998), Wester et al. (2001), and Yuan and Liberman (2009, 2011a,b) involves forced alignment systems, which, utilizing automatic speech recognition (ASR) techniques, are typically used to transform an orthographic representation of speech to a time-aligned phone-level representation. While forced alignment was not designed initially with the goal of determining which of different pronunciation variants was produced by a speaker, its underlying algorithms provide key mechanisms for such purposes. The second broad approach is the use of machine learning classification procedures, which are designed to learn patterns in data and associate those patterns with classes of objects or outcomes. In purely computational terms, the coding of many pronunciation variables is a rather straightforward classification task for machine learning. Given some acoustic information along with a set of "gold standard" data for which the correct classification is known, a supervised machine learning algorithm can extract patterns of association in the acoustic data to determine likely groupings that align with the categories. These classifying models can then be applied to new data to make predictions about the category membership of those instances.

Hand-coded data by trained analysts has often been viewed as the "gold standard" on which machine learning methods should be trained and subsequently the standards by which proficiency of different models is determined. Yet, surprisingly, the field knows little about how human analysts compare to one another in the first place. A major issue in the automated coding of sociolinguistic variables is that, given the continuous nature of production and the context-dependent nature of perception, the ground truth of whether a given token was realized as one variant or the other is often not straightforward, even for human analysts.

With a few notable exceptions (e.g., Kessens et al., 1998; Hall-Lew and Fix, 2012), very little work has actually empirically tested the extent to which different human analysts agree in their coding. Further, the approaches by Yuan and Liberman (2009) and McLarty et al. (2019) have raised the possibility that the need for human coded data training data can be avoided altogether, by taking advantage of other phonological patterns available in language data from outside the variable context. This raises questions about the necessity of human coded training data vs. ways of harnessing properties of other "variable-adjacent" data for training purposes and the resulting performance of models.

Our paper is motivated by the fact that a wide range of machine learning algorithms are now available that excel at tasks relevant to automatic coding of speech features. Yet, for the successful computational automation of the coding of sociolinguistic variables, several important questions remain outstanding before any widespread adoption can take place. For instance, for any particular situation, what is the most appropriate, or most successful, automated approach of the many available? Further, for supervised approaches, what are the most appropriate training data to lead to successful performance? What hand-labeled data are sufficient as the "gold standard" training data? And, perhaps most importantly, on what basis should the algorithm's performance be assessed? What counts as "successful," and by what metric? A growing set of techniques have been developed that would seem appropriate for automatically coding variables, and thus far different approaches have been used and with different types of training data, but, rarely has the performance of different approaches been compared to one another for the same dataset. Our paper directly takes up these questions.

We investigate a set of manual and automated sociolinguistic variable coding procedures, considering the performance (inter-analyst agreement for human coders and accuracy and signal detection performance metrics for automated procedures) and outcomes (resultant statistical patterns in variationist analyses) of human coded data and computationally coded data. We implement a series of automated coding procedures following up on techniques and suggestions in recent literature and investigate the influence of different approaches to training data on the outcomes of the procedures.

Our investigation focuses on the English sociolinguistic variable (ING), the alternation of forms like *talking* with *talkin'*. (ING) has been a central variable of interest in sociolinguistics and has fueled a wide range of theoretical and methodological advances over the past half-century. (ING) has remained a feature coded by hand in sociolinguistic research and represents an important test case for automatic variable coding because it does not have well-documented acoustic parameters that correspond with its perceived realization. That said, it is also one of the few sociolinguistic variables that has previously been addressed through automated coding techniques, with Yuan and Liberman's (2011b) study showing promise for the use of forced alignment-based automatic coding methods. For our investigation, we use data from the public Corpus of Regional African American Language (CORAAL; Kendall and Farrington, 2020a). CORAAL provides a large amount of spontaneous speech material for the development and testing of analytic methods and provides data that we can share with this paper. Additional datasets derived from CORAAL (including our hand-coded data) as well as processing scripts are available as **Supplementary Material** to this paper.

The rest of the paper is organized as follows. In the section Background, we provide further background on (ING) and on the manual and automatic coding of pronunciation features in current sociolinguistic work. We then provide more information about our data in the section CORAAL and its (ING) Data. The section Manual Coding of (ING) in the CORAAL Data describes our hand-coding procedures and the results of inter-analyst agreement assessments, which provide important baseline information for manual coding of sociolinguistic variables generally and the characteristics of our training and test data for assessing automatic coding procedures. The section Coding via Forced Alignment presents a forced alignment-based approach to automatically coding (ING) and its results. The section Coding via Machine Learning presents a series of machine learning approaches to automatic coding for (ING) and their results. Finally, the Discussion and Conclusion offers some concluding observations.

BACKGROUND

English Variable (ING)

The variants of variable (ING) are primarily described in sociolinguistic work in terms of the realization of the final nasal segment, as alveolar [n] or velar [ŋ]. Occasionally work has also considered variation in the vowel realization or other consonantal realizations, such as oral releases [ŋg] (see e.g., Kendall and Thomas, 2019), however following the majority of work we treat variable (ING) as falling into two primary pronunciation variants, which we describe as *-ing* and *-in*. While variation in (ING) is realized phonologically and occurs across different morphological forms (i.e., both within individual morphemes (*-ing*) and within larger word forms (e.g., *something*, *during*), the variable has its roots in the morphology of Old English, arising from competition between the historical present participle morpheme *-ende* and the historical verbal noun form *-ung* (Houston, 1985; Tagliamonte, 2004). Importantly for our present purposes, monosyllabic words (e.g., *thing*, *king*) are not variable and therefore not considered a part of the variable (ING). A number of papers provide extensive discussions of (ING) and its history; readers are encouraged to refer to Hazen (2008) or Kendall and Thomas (2019) for more general background.

A range of linguistic factors are known to influence (ING) realizations, including the grammatical category of the (ING) word, with verbal words (e.g., *talking*, *walking*) more likely to occur with *-in* than nouns and adjectives (e.g., *building*, *amusing*) (Labov, 1989; Tagliamonte, 2004; Hazen, 2008). Phonological context (proceeding and following environment) effects have been found in some studies but not others (Labov, 2001; Kendall and Thomas, 2019). Word frequency (Forrest, 2017), and other word characteristics (e.g., is the word "learned" or "everyday"; Tagliamonte, 2004), have also been found to play a role in (ING) realizations, although relatively few studies have examined such questions in depth. Additionally, lexical stress patterns, coupled with word length, also play a role in patterns for (ING), so two syllable words have been found to be much more likely to be realized with *-in* than longer words (Kendall, 2013).

Social factors are also known to play a role in patterns of (ING) realization. Many studies find greater use of *-in* by male speakers than female speakers (Labov, 1966; Kendall and Thomas, 2019), and social stratification is the norm, with speakers in lower social class groups using much higher rates of *-in* than speakers in high social class groups (Labov, 1966; Trudgill, 1974; Tagliamonte, 2004). Stylistic factors, such as formality and identity construction, are also known to play a role in (ING) realizations (e.g., Trudgill, 1974; Eckert, 2008; Kendall, 2013), although such within-speaker factors are outside the scope of the present investigation. While (ING) variation is ubiquitous across English varieties, speakers of African American Language (AAL), the variety sampled in our data, generally have high rates of *-in* use (Labov, 1966).

Determining the Realization of Pronunciation Variables

As described in our introduction, the growth of sociophonetics as a research area has represented an embrace of instrumental techniques for the analysis of pronunciation variation, but impressionistic coding by trained analysts remains the norm for certain variables. Sometimes impressionistic coding is done with "acoustic guidance" (e.g., by consulting spectrograms of tokens during coding) but the principal technique ultimately involves a human analyst making a categorical, auditory judgment about the variable, such as whether an instance of (ING) should be coded as *-in* or *-ing*. This manual, impressionistic analysis has remained a robust and valuable approach for analyzing variation and many consistent patterns have been identified through such data. This paper does not argue against such data, though here we make a couple of observations about their limits.

First, manual analysis of sociolinguistic variables is slow and necessarily small-scale. Transitions to bigger data and large-scale analysis in sociolinguistics are hampered by a reliance on hand-coded data. For instance, Wolfram's (1969) study of AAL in Detroit, MI – still representing one of the largest sociolinguistic community studies undertaken – quantitatively analyzed just 60 of the 728 individuals interviewed in the community (Shuy et al., 1968).

Further, despite some detailed investigations into inter-analyst patterns in related areas of phonetic transcription (e.g., Shriberg and Lof, 1991; Cucchiarini, 1993, 1996), there have been limited investigations of inter-analyst agreement in the coding of sociolinguistic variables (as well as in the acoustic measurement of sociophonetic variables; however cf. Duckworth et al., 2011). The limited studies indicate that inter-analyst agreement rates are often not high, sometimes with aspects of analysts' backgrounds playing a role in their impressionistic determinations for a variable, despite their amount of experience or training. For example, Yaeger-Dror et al. (2009) conducted a study of trained analysts' perceptions of post-vocalic /r/ realizations and found that the analysts' own dialect background influenced their judgments. And, Hall-Lew and Fix (2012) found that different professional linguists applied different thresholds for categorizing /l/ vocalization. Further, and not surprisingly, tokens that were acoustically in-between category norms were the most disagreed upon.

It is valuable to recognize that the task of impressionistic, auditory coding is in fact a kind of (often poorly-controlled) perception task, with an N of one or perhaps a few, albeit with participants (coders) that tend to be highly trained rather than naïve to the variable. Thus, we offer that it is not surprising that analysts' codes are affected by factors known to affect linguistic perception more broadly, like perceptual sensitivity, language background, or the token's context and perceiver's expectations, and that analysts' training seeks to (but does not always) eliminate such biases. For example, in studies of (ING), it has been demonstrated that naïve listeners who were asked to classify -*in*/-*ing* variation reported hearing *-in* more often in grammatical contexts where it is probabilistically more expected in English (Vaughn and Kendall, 2018), and that *-in/-ing* categorization is also affected by listeners' language background (Yuan and Liberman, 2011b).

This raises a question that is often glossed over in sociolinguistics. On the one hand, instrumental methods that measure acoustics can be easy to implement, and do not introduce the kind of bias inherent in relying on individual listeners' judgments, *but,* they gloss over the relevance of the acoustic details to listeners' auditory perception, which is arguably an important component of language in use (see Kendall and Vaughn, 2020). On the other hand, hand-coding methods that rely on coders' auditory perception reflect the reality of the perceptual system's biases, but are harder and slower to implement, and also to replicate. Thus, it is more difficult than it seems to develop and validate automated methods of impressionistic coding: What standards do we, as a field, think are important in assessing whether the system has "done a good job"? That they perform consistently (in comparison to what, human coders)? That they perform in a similar way to human coders (have a harder time with the kind of tokens that humans do)? That they would result in similar macro-level patterns across the speakers sampled (that social and linguistic factors would pattern in an expected way)? We consider these and other points in our assessments throughout this paper.

Automated Approaches to Coding Pronunciation Variables

That manual analysis is limiting for the growth of sociolinguistic studies is not a new observation and, as mentioned earlier, a handful of studies have applied computational techniques to the domains of traditional by-hand analyses. Across approaches, the basic premise is that a computational model of some kind (whether a machine learning classifier or as a part of a larger ASR workflow within a forced alignment system) learns to differentiate categories based on some source materials (*training data*) and then that model can be applied to new instances of the feature of interest (*test data*).

We focus our consideration on the use of automated coding for specifically sociolinguistic purposes, but we note that this domain of work falls within a larger area of research on the automation and validation of phonetic transcription and speech technologies like forced alignment. Much of this work has not been picked up by sociolinguistic researchers. However, it offers much to the advancement of sociolinguistic methods for both manual analyses (e.g., considerations of agreement in phonetic transcription; Shriberg and Lof, 1991; Cucchiarini, 1996) and automated approaches (e.g., considerations of how automated phonetic transcription systems perform in comparison to human analysts; Wester et al., 2001; Kessens et al., 2003; Binnenpoorte, 2006; Van Bael et al., 2007).

Using forced alignment as a tool for coding variables was one of the first applications of computational methods for automated sociolinguistic variable coding. These approaches rely on the forced alignment's system to differentiate categorical phonetic forms from acoustic information available in the signal. Yuan and Liberman, the creators of one of the first widely used forced alignment tools, Penn Phonetics Lab Forced Aligner (P2FA; Yuan and Liberman, 2008), trained their forced alignment algorithm to differentiate light /l/ and dark /l/ realizations in recordings of oral arguments from the Supreme Court of the United States (Yuan and Liberman, 2009, 2011a). In this study they took advantage of English phonological processes, whereby light and dark /l/ realizations are unambiguous in word initial (light /l/) and word final (dark /l/) position. They then trained their system on a phonological mapping of two phones: L1, light /l/ based on word initial position, and L2, dark /l/ based on word final position. After training on canonical dark and light /l/, the model was then applied to ambiguous tokens (word medial /l/) to assign one of the two labels, L1 or L2, to individual tokens. This innovated a creative solution to one of the hardest issues in automated coding, which is establishing the training data, here based on non-variable canonical representations of the phones.

In a second study – the most direct analog to the focus of the present paper – Yuan and Liberman (2011b) used a similar technique to analyze (ING) realizations in two corpora, adding a supervised learning step where their acoustic models were trained on human-labeled forms of *-in* and *-ing* for (ING) words from the Buckeye Corpus (Pitt et al., 2007). They then tested categorization on a new set of unseen data balanced for *-in* and *-ing* forms. Comparing their system's overall agreement against eight native English speakers' and 10 native Mandarin speakers' agreement across 200 tokens, they found that their approach reliably categorized *-in* and *-ing* with agreement rates comparable to agreement between native English-speaking coders (an average of 85% agreement).

Bailey (2016) extended this kind of work, testing the FAVE-Align (Rosenfelder et al., 2014) system on three variables, *t/d*-deletion, *th*-fronting, and *h*-dropping. Diverging from Yuan and Liberman's work, this study did not explicitly train a new acoustic model on the variable pronunciations or speakers, and instead aligned its British English speech with an American English acoustic model (a typical practice in forced alignment), adding alternative pronunciations for the variables to the dictionary for alignment. The system's outcomes agreed rather well with manual variable codes for *h*-dropping (∼85%) and *th*-fronting (∼81%) and less well for *t/d*-deletion (∼71%) especially in cases of *t/d* presence where inter-analyst agreement was also lower. Despite the less customized training and testing, Bailey's work again demonstrates that forced alignment categorization has overall high levels of agreement with human analysts across variables. However, Bailey also observed that FAVE-Align was sensitive to factors that human analysts were not, with FAVE accuracy decreasing as speech rate increased, while human analysts remained unaffected (though this may be the result of an acoustic model trained on a different variety).

McLarty et al. (2019) used similar reasoning to Yuan and Liberman (2009) to consider whether post-vocalic /r/ realizations could be automatically coded from a model trained on canonical, i.e., non-variable, "adjacent" contexts. Their study used CORAAL, the same public dataset as used in the present study, adopting a more standard approach to supervised machine learning, the use of support vector machine (SVM) models. In this study, McLarty et al. (2019) extracted mel-frequency cepstral coefficients (MFCCs, more on these below) at three time points across three phonological categories: vowels, pre-vocalic /r/ (which is non-variable but acoustically different from post-vocalic /r/), and post-vocalic /r/. They then trained an SVM on oral vowels and pre-vocalic /r/ tokens, and tested classification on post-vocalic /r/ and unseen vowels. The use of the non-variable phones in training was meant to provide an unambiguous representation of mappings between acoustic information and phone categories. They demonstrate overall that the results from an SVM approach applied to a social analysis of variability in CORAAL largely align with previous studies of *r*-lessness in AAL, suggesting that SVMs and the use of "variable-adjacent" phones for training may be a fruitful method for automated data coding.

Most recently, Villarreal et al. (2020) used random forests to classify post-vocalic /r/ and medial /t/ variables in New Zealand English. Unlike McLarty et al. (2019) this study relied on hand-coded tokens as the training data, with 180 acoustic measures for post-vocalic /r/ tokens and 113 acoustic measures for medial /t/ tokens. In addition to finding a good fit between their models and their training data for /r/ and for binary classification of /t/, they show that the output of their classifier predicted the ratings of trained human listeners for new tokens of post-vocalic /r/, both in terms of gradient judgment and binary classification (absent vs. present). In their paper, Villarreal et al. (2020) presented a critical assessment of McLarty et al.'s (2019) approach to training data, questioning the premise that the study's use of oral vowels and pre-vocalic /r/ tokens provided adequate acoustic information for a post-vocalic /r/ classifier and arguing against the use of such extra-variable forms as training data. While their critique raises valuable points about the need for further testing, their comments appear to miss the possible value of such an approach: Training a classifier on pronunciations outside the variable context has the potential to act as a crucial workaround for the key step in any automatic coding algorithm, which is the need for ample and robust training data. Our takeaway from their critique is that the potential use of different kinds of training data need to be tested, validated and strengthened, and on a per-variable and per-context basis, rather than assuming that one approach is inherently flawed.

Our investigation focuses on several key questions that build on these prior foundations in the automated (and manual) coding of sociolinguistic variables. But, before proceeding, we note that our study does not consider all of the important issues. For instance, we consider an unsupervised approach to coding using a state-of-the-art forced alignment system along with a set of supervised machine learning classifiers. However, we do not set up those approaches to compare them fully in an "apples to apples" way. Rather, we implement each in what we believe are typical use-case ways, embracing the rich acoustic model that the aligner is capable of building for our investigation of forced alignment (in section Coding via Forced Alignment). For our machine learning classifiers (in section Coding via Machine Learning), we focus on a set of simpler, mel-frequency cepstral coefficients (MFCCs) as the acoustic measures, without extensive parameterization or transformation. The use of MFCCs are standard in many areas of speech technology including ASR and are known to provide good representation of the acoustic signal for such purposes (Davis and Mermelstein, 1980; Huang et al., 2001). MFCCs represent extracted values (coefficients) from a mel-frequency cepstrum, which, simply put, is a non-linear spectrum of a spectrum. For variable (ING), a feature without standard acoustic measures, we believe that MFCCs are a useful acoustic representation, but we also acknowledge that further testing – into both other potential acoustic measures and the parameters for the MFCC extraction – would be beneficial. Additionally, while many of the previous studies emphasize the role of gradience in assigning values to sociolinguistic variables through the use of probability estimates of token classification (Yuan and Liberman, 2011b; McLarty et al., 2019; Villarreal et al., 2020), we limit our investigations to binary classification of (ING) to assess the general utility of different automated methods.

CORAAL AND ITS (ING) DATA

The data for this project come from the Washington DC components of the public Corpus of Regional African American Language (CORAAL; https://oraal.uoregon.edu/coraal/; Kendall and Farrington, 2020a). CORAAL is a collection of sociolinguistic interview recordings, along with time-aligned orthographic transcription, from a range of community studies focusing on African American Language (AAL), arranged into several components (subcorpora). Two of the main components are from Washington DC and these are the source of data for the present study. CORAAL:DCA contains sociolinguistic interviews from Fasold's (1972) foundational study of AAL in Washington DC recorded in 1968 (Kendall et al., 2018a). CORAAL:DCB contains sociolinguistic interviews conducted during fieldwork led by Minnie Quartey specifically for CORAAL in 2015–2018 (Kendall et al., 2018b). Both CORAAL components include extensive demographic information about the speakers, including their age, gender, and assignment to one of three socioeconomic classes [SECs: 1 (lowest) to 3 (highest)]. The two components, recorded about 50 years apart from one another, reflect some differences in sociolinguistic interview recordings, in terms of both content and recording technology. They also can be expected to involve recordings with different acoustic properties (the DCA interviews were recorded on reel-to-reel tape and digitized in ∼2013; the DCB interviews were recorded digitally using modern solid-state recording hardware; see Kendall and Farrington, 2020b). Our investigation uses both sets of recordings together, and thus provides baseline performance information for the classification of tokens from somewhat heterogenous data. For sake of space, we leave considerations of differences between the two components for future work. It should be noted that our paper does not focus on AAL, but all of the speakers examined identify as Black/African American.

(ING) variation in the CORAAL data was the focus of a (2019) paper by Forrest and Wolfram, who used a set of speakers available in an early version of CORAAL to explore this variable. They focused on speakers in age groups 2–4, with a goal of achieving balance across demographic categories. While our data are independent of the tokens impressionistically examined in that work, their paper provides a preliminary view of the patterns in CORAAL. They identified socioeconomic differences in the rates of (ING) variation in both components of CORAAL, with high rates of *-in* use among the lowest SEC group (above 93% in DCB) and decreasing rates among the higher SEC groups, along with an interaction between gender and SEC for DCA, where males used much higher rates of *-in* than females in the lower SEC groups. Grammatical conditioning has been found for (ING) in several varieties of English (Tagliamonte, 2004; Hazen, 2008), where the *-in* variant is more likely in verbs than in forms like nouns and adjectives. In DCA and DCB, Forrest and Wolfram found only weak grammatical effects, although verbs did exhibit the highest rates of *-in* in both components, aligning with other work on (ING). While our data source is the same, we would not necessarily anticipate identical results to Forrest and Wolfram (2019) for methodological reasons. In our study, we included speakers from a wider range of age groups and also extracted our (ING) tokens to code a random sample from all available (ING) tokens of the speakers selected (e.g., we did not implement type/token limits), rather than the sequential, systematic token inclusion procedures typically used in sociolinguistic analyses.

To examine (ING) in CORAAL, we mined the DCA and DCB components for data. All speaker turns containing non-monosyllabic words with word-final "ing" were extracted from the publicly available R version of the corpus text for DCA and DCB. Interviewers from DCA, who for the most part were not African American, and a few tokens from "miscellaneous" speakers, were removed from the dataset. We also extracted words from a separate, phone-level aligned version of the transcripts, generated using the Montreal Forced Aligner (MFA; McAuliffe et al., 2017); this process is described further in the section Coding via Forced Alignment. We merged these two versions of CORAAL to select tokens of (ING) for analysis. In addition to the variable (ING) words extracted from the corpus, words with word final [ɪn] and [ɪŋ] that are not in the variable context for (ING) (e.g., *in*, *thin*, *Chaplin*, *vitamin* for [ɪn] and monosyllabic *-ing* words, like *thing*, *bring*, *cling*, *wing*, for [ɪŋ]) were also extracted for comparison with the variable (ING) cases.

For each variable (ING) word and each non-variable *IN* and *ING* word, 12 MFCCs were extracted from four temporal measurement points in each final vowel+nasal portion of the word, 25, 50, 70, and 90% of the vowel+nasal segments' combined duration, following prior work citing the importance of vowel quality in (ING) classification (Yuan and Liberman, 2011b). These were based on the segment alignments from the MFA forced alignment. The MFCCs were extracted using the tuneR package in R (Ligges et al., 2018). Words with final vowel+nasal segments that were ≤50 milliseconds or for which our MFCC extraction process otherwise failed to obtain MFCCs were dropped from the dataset. This left a total of 8,255 *IN* words and 1,436 *ING* words in the non-variable MFCC data and 12,041 (ING) words in the variable data. Preliminary tests assessed a range of different MFCC extraction parameters and their impacts on the later classification steps of our process but we found little impact of minor changes to the MFCC parameters. We do not focus on testing different MFCC time points or window lengths in this paper but our initial investigations indicated that four temporal measurement points for the extraction of MFCCs performed better than tests with two or three time points, even though fewer time points allowed for the inclusion of shorter vowel+nasal segments (so led to an increase in the total number of tokens that could be considered. Data sources and R code along with more information about our procedures, including the specific settings used for e.g., MFCC extraction, are provided as **Supplementary Material**).

MANUAL CODING OF (ING) IN THE CORAAL DATA

Before considering the ability of automated, computational approaches to code instances of (ING), it is important to assess the nature of such data from the perspective of human coders. As discussed earlier (in the section Determining the Realization of Pronunciation Variables), very little work in sociolinguistics has published accounts of inter-analyst agreement in the coding of variables (cf. Hall-Lew and Fix, 2012). Understanding the degree to which human coders agree about codes for (ING) is important before we can assess the performance of machine coding of the variable. Further, human annotations for gold standard training and test data are a major component of most machine learning classification approaches, so understanding the properties of the human coded data is important for the other steps of our research project.

For the human coded data, 50 tokens were randomly subsampled per speaker from the larger dataset, for 24 speakers. All of the speakers are African American and were selected to include the major demographic categories included in CORAAL's sampling – speaker gender, age, and socioeconomic status – but with an emphasis on the lower SEC groups. **Tables 1A,B** display the breakdown of speakers. In addition to the 1,200 tokens sampled from these 24 speakers, 100 tokens were randomly subsampled from the main interviewer in the DCB corpus, an African American female in her 30s. This interviewer is by far the speaker with the most recorded speech in CORAAL and we thought including a sample of (ING) data from her speech would be useful.

TABLE 1A | Speakers included in Dataset B from CORAAL:DCA.

	Socioeconomic group 1	Socioeconomic group 2 and 3
Age group 1 (<19)	DCA_se1_ag1_f_04 (95.8%) DCA_se1_ag1_m_07 (95.8%)	DCA_se2_ag1_f_02 (69.8%) DCA_se2_ag1_m_05 (37.5%)
Age group 2 (20–29)	-	DCA_se3_ag2_f_02* (6.0%)
Age group 3 (30–50)	DCA_se1_ag3_f_02* (34.7%) DCA_se1_ag3_m_01* (89.1%)	DCA_se2_ag3_m_01* (87.8%)
Age group 4 (>51)	-	-

In parentheses is the percentage use of -in by the speaker based on their 50 tokens in Dataset B.
**Also included in Forrest and Wolfram (2019) analysis.*

TABLE 1B | Speakers included in Dataset A (in *gray italic* font) and Dataset B (plain font) from CORAAL:DCB.

	Socioeconomic group 1	Socioeconomic group 2
Age group 1 (<19)	DCB_se1_ag1_f_03 (77.1%) DCB_se1_ag1_m_02 (89.1%)	DCB_se2_ag1_f_01 (83.3%) DCB_se2_ag1_m_01 (83.7%)
Age group 2 (20 to 29)	*DCB_se1_ag2_f_02* (84.6%)* *DCB_se1_ag2_m_01* (100%)*	DCB_se2_ag2_f_02* (10.0%) DCB_se2_ag2_m_01* (87.2%)
Age group 3 (30 to 50)	DCB_se1_ag3_f_03 (93.9%) DCB_se1_ag3_m_02* (88.0%)	DCB_se2_ag3_f_02 (62.0%) DCB_se2_ag3_m_02* (60.4%)
Age group 4 (>51)	DCB_se1_ag4_f_01 (97.9%) DCB_se1_ag4_m_01 (84.0%)	DCB_se2_ag4_f_05 (83.3%) DCB_se2_ag4_m_01 (94.7%)

Plus DCB_int_01 (female interviewer, mid 30s; -in: 58.7%).
In parentheses is the percentage use of -in by the speaker based on their 50 tokens in Dataset A or B.
**Also included in Forrest and Wolfram (2019) analysis.*

For two of the speakers, DCB_se1_ag2_f_02 and DCB_se1_ag2_m_01 (both in the lowest socioeconomic group and in the 20–29 age group), all seven authors coded each of the tokens. We hereafter refer to this as Dataset A, and we use it to assess inter-analyst agreement patterns for a(n albeit small) dataset coded by more than just a few analysts. For the other 22 speakers and the interviewer, three analysts coded each token. We hereafter refer to this as Dataset B. Thus, for Dataset A we have seven independent ratings for 100 of the (ING) cases and, for Dataset B, three independent ratings for the other 1,200 tokens.

In addition to the hand-coded tokens just described, an additional set of 900 tokens, hereafter Dataset C, were randomly selected from CORAAL:DCA and CORAAL:DCB with no sampling criteria other than that these tokens did not come from interviewers in DCA (who, again, were generally not speakers of AAL) and that did not overlap with the 1,300 tokens sampled for the Datasets A and B. Dataset C includes tokens from 113 speakers, with an average of 8.0 tokens per speaker and a standard deviation of 8.4 (a maximum of 69 for the main DCB interviewer, who is the person with the most speech in the corpus, to a minimum of 1 token each for 11 speakers, who generally are speakers who contribute only small amounts of speech to the corpus). This final set of tokens was coded by two of the authors and is used in some of our analyses as an additional test dataset. We note that as a random sample of the entirety of CORAAL:DCA and DCB, Dataset C is useful for examining the overall patterns that might occur across the complete dataset. It also allows us to test samples of speech from speakers who are not present in any of the data we use for training models. We also note, however, that Dataset C is somewhat artificial as an example of a sociolinguistic dataset, since most sociolinguistic studies will sample speakers in more systematic ways and will not, for example, develop a dataset with such imbalanced tokens across speakers. Nonetheless, we believe that Dataset C provides us additional value as a test case for our automated techniques.

In order to code the tokens, the human analysts worked from spreadsheets of excerpts from orthographic transcriptions, with each excerpt line containing one specified (ING) word. Each line contained a direct link to the audio for the token's utterance via the online interface to the corpus (http://lingtools.uoregon.edu/coraal/explorer/browse.php). Analysts were instructed to listen to the token in context, and code the (ING) tokens auditorily according to the following categories: "G" if the form was clearly *-ing*, "G?" for cases where the analyst believed it was *-ing* but wanted to register a lack of confidence, "N" if the form was clearly *-in*, and "N?" for *-in* but without confidence. Finally, analysts were instructed to use "DC," for *don't count*, if for some reason the token did not appear to be a good candidate for analysis. There are several reasons a token could be a *don't count* form, ranging from instances where our initial extraction selected tokens that simply were not good for analysis (e.g., the speech overlapped with other simultaneous speech in the recording or the token involved some disfluency on the part of the speaker) to cases where the form was determined to be too unclear to code. Since the (ING) tokens were selected from the corpus by script, the coders were instructed to use DC codes as liberally as necessary and we might expect a higher number of DC cases here than in typical variationist analyses which pre-select tokens for inclusion using more deliberate processes. Aside from these reasons for marking a token as DC, all non-monosyllabic *ing*-final words were included as candidates for the (ING) variable. We note that researchers examining (ING) have implemented different practices regarding some aspects of the variable, such as whether lexical exclusions apply (e.g., excluding words like *anything* and *everything* which tend to favor *-ing* or words like *fucking* which tend to favor *-in*). Our practices follow Hazen (2008) and Kendall and Thomas (2019) in not applying any such exclusions (see also sections English Variable (ING) and Automated Approaches to Coding Pronunciation Variables).

Importantly, we note that all of the authors are trained linguists with varying degrees of research experience with AAL, however none are speakers of AAL. Research experience of the authors ranges from extensive transcription of interviews in CORAAL to research and publications on AAL more broadly. This fact may be one potential factor affecting our coding, as language backgrounds have been observed to influence perceptual categorization of variants. We note, however, that this fact – non-AAL speakers coding AAL data – is not unusual in sociolinguistic studies, so may be representative of a more widespread limitation of impressionistic coding in sociolinguistics. The question of language variety background and inter-analyst agreement in sociolinguistics is outside the scope of our paper, but warrants further attention.

Dataset A: Inter-analyst Agreement Among Seven Human Coders

We begin by considering the patterns of agreement in Dataset A, the 100 tokens coded by all seven analysts. This is admittedly a small dataset but little sociolinguistic work (or other linguistic annotation description) has reported coding outcomes by more than a few analysts, so we begin by assessing what kinds of agreement coding might yield across all seven analysts.

Of the 100 tokens coded by all of the analysts, 20 tokens received at least 1 DC (*don't count*) code and 9 of the tokens (5 of which overlapped with tokens that also received DC codes) received at least one low confidence (N? or G?) code. In order to simplify the treatment here (i.e., for sake of space), we collapse over the low confidence codes (so N and N? are collapsed to *-in* here and G and G? are collapsed to *-ing*). The breakdown of these codes for the 100 tokens are displayed in **Table 2**. The high number of forms coded as *don't count* (20% of the data received at least one such vote) is likely a function of the instructions to use DC liberally in order to catch erroneous tokens that were selected by our automated selection procedure (e.g., cases of speaker overlap). Six tokens received 3 or 4 DC votes, which likely indicate that those tokens should indeed be discounted from an analysis, but 10 tokens received only 1 DC vote, which suggests that our DC criteria could have been clearer to the coders. One take-away from the DC forms alone is that subjective decisions about coding involve not only coders' impressions of what form they perceive but also what constitutes a "countable" instance of the variable in the first place.

Fifty-eight tokens were coded as *-in* by all seven analysts. An additional 21 tokens were coded by six of the seven analysts as *-in* (with 12 coded with one *-ing* and the other 9 coded by one analyst as DC). Only three tokens were coded by all analysts as *-ing*, which we take as evidence of the low rate of use of *-ing* by these two working class speakers rather than as something inherent about coding *-ing* cases as opposed to *-in* cases. Most of the other possible outcomes occurred in this small amount of data, with, for instance, one token being coded by four analysts as *-ing* and three analysts as *-in*. Overall, a measure of inter-analyst agreement using Fleiss' Kappa for multiple raters (Conger, 1980) yields a $k = 0.39$ with significantly better agreement than chance for each of the three categories (*-in*: $k = 0.38$, *-ing*: $k = 0.52$; DC: $k = 0.22$). However, the agreement values still fall only in the "fair" to "moderate" agreement range according to many assessments of inter-analyst agreement (Landis and Koch, 1977). Removing the DC cases, a clear source of disagreement among the analysts for the tokens in Dataset A, improves the agreement rates substantially to $k = 0.54$. This small sample coded by many analysts demonstrates that we need to expect some amount

TABLE 2 | (ING) codes for the 100 tokens in Dataset A coded by seven analysts (=how many analysts coded the tokens using a particular code?).

	0 analysts	1	2	3	4	5	6	7 analysts
-in	3	1	2	3	4	8	21	58
-ing	76	13	5	0	1	1	1	3
DC	80	10	4	3	3	0	0	0

TABLE 3 | (ING) codes for 1,135 tokens in Dataset B coded by three analysts (not including tokens with DC codes).

Codes:	N-N-N	N-N-N?	N-N?-N?	G?-N-N	G-N-N	G-N-N?	G-G?-N	G-G-N	G-G-G?	G-G-G
N:	697	3	1	2	107	2	5	39	8	271
%:	61.4%	0.3%	0.1%	0.2%	9.4%	0.2%	0.4%	3.4%	0.7%	23.9%
	Agree *-in*: 701 (61.8%)			Disagree: 155 (13.7%)					Agree *-ing*: 279 (24.6%)	

of disagreement as normal in manually coded pronunciation variables like (ING).

As reported in **Table 1** earlier, the two speakers included in Dataset A were heavy users of *-in*. Removing all tokens which received any DC votes and taking a majority-rules view of the realization – where we take the majority of analysts' codes as the category for a token – only six tokens would be assigned as *-ing* across the two speakers and all were produced by the female speaker, DCB_se1_ag2_f_02 (*-in* rate = 84.6%). The male speaker, DCB_se1_ag2_m_01, had categorical use of *-in*. In retrospect, it would have been more useful to include speakers who were more variable in Dataset A, but we did not, of course, know their rates of use before selecting the speakers for inclusion.

Dataset B: Inter-analyst Agreement Among Three Human Coders

For further consideration we move to assess the codes generated by three analysts for the other 22 speakers and the interviewer. To do this, we first removed all tokens that were coded as DC by any of the analysts. This removed 65 tokens from the 1,200 tokens coded by three analysts, leaving 1,135 tokens. The breakdown of codes is presented in **Table 3**. An assessment of the inter-analyst agreement using Fleiss' Kappa yields $k = 0.77$ for the data including the low confidence ratings (N, N?, G?, and G) and $k = 0.79$ if the confidence codes are collapsed (i.e. just assessing *-in* vs. *-ing*). These are high levels of agreement, in the "substantial agreement" range by common rules of thumb. In simpler terms, and collapsing the confidence marks, the coders agree (all three assign the same major code) for 980 tokens (86.3% of the 1,135 tokens).

Taking a majority-rules view of the coded data – i.e., any tokens with two or more G or G? codes count as *-ing* and two or more N or N? codes count as *-in* – suggests that, overall, the speakers produced 812 (71.5%) of the tokens as *-in* and 323 (28.5%) as *-ing*. These values provide both a useful benchmark for the potential results of automated approaches to coding CORAAL's (ING) data. They also provide a useful starting place for training data for such a coding system. We use Dataset B extensively for training and testing automatic coding routines in the section Coding via Machine Learning.

TABLE 4 | (ING) codes for 900 tokens in Dataset C coded by two analysts.

Codes:	N-N (Agree *-in*)	N-G (Disagree)	G-G (Agree *-ing*)
N:	569	104	227
%:	63.2%	11.6%	25.2%

Dataset C: Inter-analyst Agreement Among Two Coders

As an additional dataset for assessing the performance of automated coding methods, two analysts coded (ING) for the additional set of 900 tokens from CORAAL. These two raters obtain 88.4% agreement for this second set, with a Cohen's $k = 0.73$. The breakdown of these tokens is presented in **Table 4**, showing overall rates of *-in* (63.2%) and *-ing* (25.2%), with 11.6% of the tokens as ambiguous, having been coded as *-in* by one analyst and *-ing* by the other. While these tokens are sampled more randomly than the sample in Dataset B, comprising a wider assortment of speakers across all of the CORAAL:DCA and CORAAL:DCB, these rates are taken as comparable to the 71.5% *-in*/28.5% *-ing* rates in Dataset B. Dataset C is used as test data in our assessments of automated coding routines in the section Coding via Machine Learning.

CODING VIA FORCED ALIGNMENT

As a first step toward automatically coding variable (ING) in CORAAL, we submitted CORAAL:DCA and DCB (v. 2018.10.08) to forced alignment, using the Montreal Forced Aligner (MFA; version 1.0). This alignment was done using MFA's train and align option, which creates an acoustic model based entirely on the dataset itself. For the pronunciation dictionary, we provided the Montreal Forced Aligner (MFA) with an edited version of the Carnegie Mellon University pronunciation dictionary that, crucially, included two pronunciation options for each (ING) word (e.g., *bringing* was represented in the

TABLE 5 | Human codes for Dataset B along with MFA's pronunciation assessment.

Human codes:	N-N-N	G-N-N	G-G-N	G-G-G	*Totals*
MFA = *-in*	633 (90.3%)	94 (84.7%)	27 (61.4%)	69 (24.7%)	*823 (72.5%)*
MFA = *-ing*	68 (9.7%)	17 (15.3%)	17 (38.6%)	210 (75.3%)	*312 (27.5%)*
Totals	*701*	*111*	*44*	*279*	

Shading indicates cells where human codes and MFA agree.

pronunciation dictionary supplied to MFA with both B R IH1 NG IH0 N and B R IH1 NG IH0 NG as potential pronunciations). These entries were added to the dictionary using a script, which is included in the **Supplemental Material**. Speaker adapted triphone training was used in the train and align option in MFA, where speaker differences and context on either side of the phone are taken into account for acoustic models.

Before proceeding, we note that the use of a large, high variability training data set (number of speakers, acoustic quality, etc.) is expected to provide a more robust acoustic model for alignment (McAuliffe et al., 2017). That is, MFA was trained on all of the acoustic information available in DCA and DCB and allowed to assign phone labels to all (ING) words, with no data held out for separate testing. This differs from the training and testing approaches we take up in the section Coding via Machine Learning, but follows typical practice for use of modern aligners like MFA (However, unlike many uses of aligners, and e.g., the approach used by Bailey (2016) to code variables, our MFA acoustic models were trained specifically on CORAAL data and in a way that allowed the model to learn different pronunciations for (ING)). We do this to emulate the standard workflow that we would expect of sociolinguistic studies using forced alignment techniques; since there is no need for hand-coding training data in this unsupervised method, there is not the same motivation for testing the forced alignment system on a held out subset of the data as is the case when using hand-coded training data in our supervised classification techniques. Therefore, we emphasize that we expect MFA to do quite well, since the test dataset is subsumed by the training dataset.

As a first assessment of the codes obtained from the MFA alignment, we consider its performance compared to the human analysts' judgments for Dataset B, with the codes from the three analysts collapsed over confidence ratings (i.e., G and G?→G, N and N?→N). This is shown in **Table 5**.

Overall, the MFA outputs yield some similarity to the human coders but in some key places differ substantially. In terms of disagreement, MFA indicated a pronunciation of *-in* for 69 (24.7%) of the cases all three human analysts agreed were *-ing* and *-ing* for 68 (9.7%) cases where all three humans coded *-in*. This diverges from the humans for 12.1% (137/1,135) of the tokens. This difference is on par with the disagreements identified among the human coders for both Datasets B (13.7%) and C (11.6%). Further, the overall rates of MFA's assessment of the pronunciations of (ING) are quite similar to those of the human coders, with MFA assigning 72.5% of the (ING) cases as *-in* to the human coders 71.5%.

An additional way to assess the relative output of the forced alignment's phone labels in comparison to human coders is to ask how the outcome of a variationist-style statistical analysis might compare between the two approaches. While the evidence indicates that about 12% of the individual tokens mismatched between MFA and human coders, are the overall patterns similar, especially for factors that sociolinguists tend to be interested in? We focus here on the social determinants of (ING), each speaker's gender, age, and SEC, along with two linguistic factors, the grammatical category of the (ING) word and the length of the word in syllables. For grammatical category, we limit our focus to a binary comparison which we refer to as verb-like (V-like) vs. noun-like (N-like) forms. These were generated based on a part-of-speech tagged version of the CORAAL being developed (Arnson et al. in progress). V-like includes all of the verbal POS tags along with the pronouns *something* and *nothing* and the words (*mother*)*fucking*, which tend to pattern like verbs in having higher rates of *-in*. N-like includes nouns and adjectives along with prepositions (e.g., *during*) and the pronouns *everything* and *anything*, which are known to have lower rates of *-in*. Word length (in syllables) was generated for each word using a script available from Kendall (2013).

Table 6 displays the results for logistic regression models of the (ING) patterns in Dataset B. Model I assesses the majority-rules view of the human coded data, where each (ING) is assigned *-in* or *-ing* based on two or more coders' agreement, with the dependent variable as *-in*. Model II assesses the MFA output for the same tokens, again with *-in* as the dependent variable. The models include random intercepts for speaker and word and test main effects (no random slopes or interactions were tested) for the three social factors and two linguistic factors just mentioned. Word length is included as a continuous predictor; the other factors are categorical and included using simple (dummy coded) contrasts. For socioeconomic status, the reference level is set to SE1, the lowest SEC group. For age group the reference level is set to the oldest speakers, age group 4 (speakers who are 51+ years old). We note that age is modeled as a categorical predictor, using the age group categories provided in CORAAL. (ING) is typically found to be a stable variable in sociolinguistic community studies, not undergoing change. However, (ING) is often found to show age-grading, with middle-aged speakers showing less use of *-in* in comparison to young and old speakers (due in part to linguistic marketplace factors) (see e.g., Wagner, 2012). While a full analysis of (ING) in CORAAL is beyond the scope of this paper, the expectation of such age-graded patterns motivates our inclusion of age as a factor and the inclusion of age through CORAAL's categorical age groups provides a simple

TABLE 6 | Logistic mixed-effect regression models for (ING) in Dataset B (1,135 tokens).

	Model I: human coders ($N = 1,135$ tokens)			Model II: forced alignment output ($N = 1,135$ tokens)		
	Est.	Std. Err.	*p*	Est.	Std. Err.	*p*
(Intercept)	4.78	1.32	0.0003***	2.24	0.82	0.0060**
Corpus (DCB, vs. DCA)	0.68	0.69	0.3278	−0.44	0.38	0.2416
Gender (male, vs. female)	1.29	0.63	0.0391*	0.88	0.34	0.0089**
AgeGrp (AG1, vs. AG4)	−1.29	1.01	0.2017	−0.23	0.53	0.6566
AgeGrp (AG2, vs. AG4)	−3.73	1.25	0.0028**	−1.80	0.66	0.0060**
AgeGrp (AG3, vs. AG4)	−1.88	0.98	0.0553.	−0.73	0.51	0.1526
SEC (SE2 or 3, vs. SE1)	−1.28	0.67	0.0568.	−0.37	0.36	0.3079
GramCat (N-like, vs. V-like)	−1.14	0.37	0.0019**	−1.08	0.31	0.0005***
Word Len (# Sylls)	−0.68	0.29	0.0189*	−0.03	0.25	0.9103

****p < 0.001, **p < 0.01, *p < 0.05, .p < 0.1.*

means to uncover non-linear age differences in the data that might be missed through a simple linear treatment of age as a continuous predictor.

There are some notable differences between the human coded and forced alignment coded data, but also a number of similarities. Models do not identify a significant difference between the two CORAAL components. Both models identify the expected difference between verb-like words and noun-like words, with noun-like words significantly disfavoring *-in.* Both models also indicate that age group 2, speakers between the ages of 20–29, are significantly less likely to produce *-in* than the oldest group of speakers. Neither model finds the other two age groups significantly different from the oldest speakers, although the age group 3 speakers (between 30 and 50) come close to a *p*-value of 0.05 in Model I. Neither model identified significance for SEC differences among the speakers, although the human coded data in Model I approach significance. Both models are also similar in identifying a significantly greater use of *-in* by male speakers. The statistical outcomes suggest that the two approaches to coding were somewhat similar in their sensitivity to social patterns in these data (While we don't focus on the substance of these patterns here, they are roughly in line with sociolinguistic expectations, e.g., with greater use of *-in* by males than females and the appearance of age-grading patterns for (ING)). One striking contrary point, however, is that the word length effect is only significant in the human coded data. The fact that the forced alignment data do not capture this statistically significant pattern in the human coded data may suggest a major difference in how human coders treat, and hear, variable (ING) in comparison to the automated alignment algorithm (see also Yuan and Liberman, 2011b; Bailey, 2016).

CODING VIA MACHINE LEARNING

We turn now to consider machine learning based approaches more directly, where the coding algorithm can be trained specifically around the features of interest. While a host of potential machine classifiers are available, we focus on the two cases that have seen recent use for sociolinguistic variable coding, support vector machines (SVMs) and random forest (RF) classifiers.

SVMs are a supervised machine learning algorithm that have seen widespread use in classification (Boser et al., 1992), as well as recent work in sociolinguistics (McLarty et al., 2019). The basic mechanism of the SVM approach involves a model identifying a hyperplane in a multidimensional feature space that best separates categories based on those features. One key piece of the SVM architecture is the ability to apply different kernel functions, which allow for different kinds of separating hyperplanes between classification categories. There are other parameters that are customized for SVM algorithms, namely the "cost" of constraints violation parameter *C* and, for radial kernels, *gamma*, which determines how much influence the model places on each training example. There is no single best method for how to parameterize an SVM classifier, with most guidance suggesting an empirical approach, determining the best parameters (so-called "tuning") based on performance for the data and the problem at hand. We used the e1071 package for R (Meyer et al., 2019) interface to the C++ libsvm implementation (Chang and Lin, 2001) for all SVM models.

RFs are an approach that have seen growing use in sociolinguistics more generally, e.g., for the analysis of sociolinguistic data (Tagliamonte and Baayen, 2012). As mentioned earlier, Villarreal et al. (2020) applied RFs for their automatic coding of sociolinguistic post-vocalic /r/ and medial /t/ data. RFs are a procedure that expand upon classification and regression trees, a common recursive partitioning method, generating many individual trees on a dataset to generate a partitioning solution that is generalizable beyond a specific set of data. RFs have fewer parameters to customize than SVMs, but still benefit from model tuning. A number of random forest implementations are available. We used the randomForest package in R (Liaw and Wiener, 2002).

Altogether, we build, tune, and test three types of machine learning classifiers, an SVM with a linear kernel (hereafter "linear SVM"), an SVM with a radial kernel ("radial SVM"), and a random forest ("RF"). For each, we use two kinds of training data. In the first case, in the section (ING) Classification, Using Human Coded Training and Test Data, we proceed through a somewhat typical supervised machine learning case, where we use subsets of the human coded data in Dataset B to train the classifiers. In the second case, the section (ING) Classification, Using Variable-Adjacent Productions as Training Data, we explore the kind of approach proposed by Yuan and Liberman (2009, 2011a) and McLarty et al. (2019), where "variable-adjacent" phonetic material, from outside the variable context, is used as training data. In both cases, we use human coded data for testing the models.

(ING) Classification, Using Human Coded Training and Test Data

We start by assessing the success of the three classifiers trained on the hand-coded data. To do this, we use a 10-fold cross validation approach, using the 1,135 tokens in Dataset B, which were manually coded by three analysts, in a series of training and testing assessments.

First, Dataset B was trimmed to exclude tokens that did not occur in our MFCC extracted data (most often because they were too short for our MFCC extraction, although in some cases tokens could not be matched due to multiple potential candidates in the same utterance). This removed a large number of tokens (29.2% of the data), leaving 803 tokens. Of these, 501 (62.4%) were coded as *-in* and 302 (37.6%) were coded as *-ing* based on the majority-rules codes for three raters. For *-in*, 429 (85.6%) of the cases were agreed upon by all three coders, with the remaining 72 cases (14.4%) having agreement by two of the three coders. For *-ing*, 268 (88.7%) of the cases were agreed upon by all three coders with the remaining 34 tokens (11.3%) having agreement by two of the coders.

Parameters were chosen for the three classifiers by a model tuning step, which conducted a grid-search over candidate settings. After determining parameters based on the entire dataset, the data were randomly divided into 10 "folds," each containing 10% of the tokens, and then each of the three classifiers was trained, using the 48 MFCCs (12 MFCCs for each of four measurement points), for 9 of these folds (90% of the available hand-coded (ING) tokens), using the majority-rules category (*-in* or *-ing*) as the correct outcome. Each classifier was then tested on the held out 10% of tokens, assessing the model's predictions against the majority-rules coding for those tokens. We repeat this over 10 iterations so that each 10% fold of the data is used as a test case with the other 90% as the training data. Cross-fold validation such as this helps to assess how stable the classifier is to its training and test data. For each iteration, we measure the model's accuracy along with other performance metrics, for both the training data (how well was the model able to fit the training data?) and the testing data (how well was the model able to predict the outcomes for previously unseen data?). We also report the overall percent predicted as *-in* by the models, which helps to show the extent to which each model is over- or under-predicting *-in* vs. *-ing*. Finally, we calculate standard signal detection measures, precision, recall, F1 score, and area under the ROC curve (AUC). These measures provide more insight than accuracy alone, and, are especially valuable since the data are not balanced across *-in* and *-ing* realizations. That is, since 62.4% of the tokens in the (trimmed) Dataset B were coded as *-in* by the human coders, a classifier that always chose *-in* would be accurate 62.4% of the time. Signal detection measures provide more robust performance measures for cases like these. Our reporting of many performance metrics is also meant to provide baseline information for future work on automated coding procedures. Accuracy information from the individual runs of the 10-fold cross validation procedure are displayed in **Figure 1** and other performance metrics from across the 10 runs are summarized in **Table 7**.

Overall, the classifiers fit the training data with accuracies of 83.0% (linear SVM), 93.6% (radial SVM), and 81.8% (RF) on average. The models' predictive accuracy, their ability to match the human codes for the testing data on each run, is also decent, matching the gold standard codes 78.2% (linear SVM), 86.1% (radial SVM), and 82.1% (RF) on average. We note that these numbers are slightly lower than the agreement between the forced alignment algorithm and the human raters, although the radial SVM's performance is close, only diverging from humans' judgments 13.9% of the time.

We include **Figure 1** as an illustration of the performance of the three classifiers, and to underscore the utility of the 10-fold cross validation method. As visible in the figure, across the iterations we see general stability in the models' fits and performance for the training data (the lines for the training data are relatively flat). And the average amount of actual *-in* use is stable across each of the training datasets. This is not surprising given the amount of training data; shifts between which specific 10% folds are excluded do not lead to large changes in the overall proportion of *-in*. Further, the overall good fit, over 93% for the radial SVM, for the training data is somewhat expected. These techniques should be able to model accurately the data provided for training. The more important question is their performance on the testing data.

The testing data, with many fewer tokens (10% of Dataset B, as opposed to the 90% in the training data), are more erratic in the actual rates of *-in* (ranging from 53.8 to 68.8%) across the folds. Predictably, the models appear sensitive to this with, generally, correspondingly variable predictions across the testing folds. The accuracy on test folds varies quite a bit, however, with the worst accuracy at 70.0% for Fold 8 of the linear SVM and the best accuracy, of 90.0%, at Fold 4 for the RF and Folds 5 and 7 for the radial SVM. While the accuracies are at times rather good, the models generally over-predict *-in* in both the training and test data (as seen in the higher lines for the models' *-in* prediction rates in comparison to the actual rates).

As an additional test of the models' performance, we trained each model (a linear SVM, a radial SVM, and an RF) on the entire set of data in Dataset B and then tested these models on the two-analyst Dataset C data. 580 of the 900 tokens of Dataset C were mapped to the extracted MFCC data; the unmapped tokens were removed. 516 of the remaining (ING) tokens had agreement by the two coders (311 *-in*, 205 *-ing*) and we focus on these tokens. The models yield slightly lower performance than they did on Dataset B, over-predicting *-in* at a higher rate than for Dataset B. This difference in performance makes some sense given that Dataset C contains a wider range of speakers, many of whom

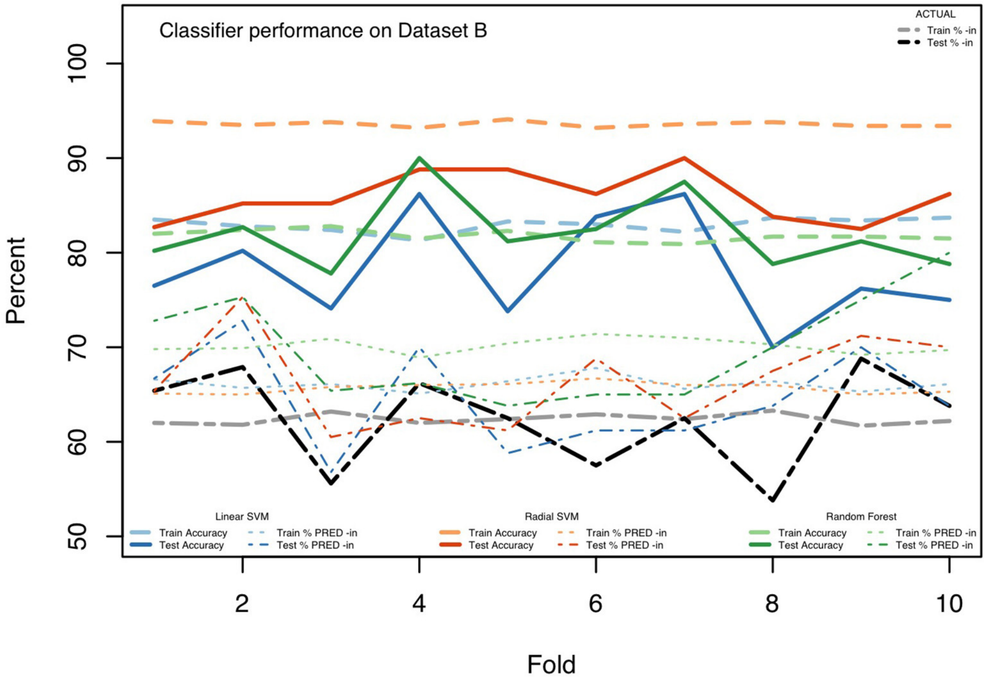


FIGURE 1 | Classifier performance on Dataset B.

TABLE 7 | Classification performance for models on Datasets B and C.

	Dataset B test performance (mean (and std. dev.) across 10 tests; training on 90%, testing on 10%) Actual *-in* rates (mean and std.dev): 62.4% (5.2)						**Dataset C performance (training on 100% Dataset B, testing on Dataset C) Actual *-in* rates: 60.3%**					
	Pred. *-in* rate	**Accuracy**	**Precision**	**Recall**	***F*1 score**	**AUC**	**Pred. *-in* rate**	**Accuracy**	**Precision**	**Recall**	***F*1 score**	**AUC**
Linear SVM	64.5% (5.3)	78.2% (5.6)	0.84 (0.05)	0.81 (0.06)	3.25 (0.23)	0.76 (0.06)	70.7%	78.7%	0.91	0.78	3.10	0.75
Radial SVM	66.6% (4.8)	86.1% (2.8)	0.92 (0.04)	0.86 (0.05)	3.46 (0.20)	0.84 (0.04)	69.0%	79.3%	0.90	0.79	3.15	0.76
RF	69.9% (5.6)	82.1% (3.9)	0.92 (0.03)	0.82 (0.06)	3.28 (0.24)	0.79 (0.05)	72.5%	77.7%	0.92	0.76	3.05	0.74

were not included in the models' training, Dataset B. Nonetheless, the models still obtain accuracies in the high 70% range.

The models all perform relatively similarly, although the radial SVM performs slightly better than the linear SVM and RF models by all measures of performance. The RF model does better than the linear SVM for Dataset B, but not quite as good as the linear SVM for Dataset C. As noted above, all models appear to over-predict *-in* rates, with the RF models doing this the most. Across the board, models' precision is slightly better than their recall.

As a final assessment, we consider the outcomes of logistic regression on the Dataset C tokens. **Table 8** provides the output of models similar to (i.e., with the same modeling structures as) those presented in **Table 6** but here presenting simple mixed-effect models for the 516 tokens of Dataset C which had MFCC measures and agreement among the two human coders. The only structural difference between the models for Dataset C and B is that for Dataset C the SEC factor was included in the model with three levels (1 lowest (reference level) to 3 highest), as annotated in CORAAL (Dataset C with a wider range of speakers includes a more complete sampling across all three SEC groups, whereas Dataset B only included limited data from SEC 3). Model III presents a model fit to the human coded Dataset C tokens, with the dependent variable being whether the humans coded the token as *-in*. Model IV presents a model fit to the same data with the dependent variable being whether the radial SVM classifier classified the token as *-in*. We present only this classifier's outcomes for space and chose it because it performed slightly better than the other two classifiers.

Model III, for the human coded data, indicates that gender and socioeconomic status are significant factors in *-in* realization for Dataset C. Males use *-in* at significantly greater rates and the highest SEC group uses *-in* at significantly lower rates than the lowest SEC group, which is the reference level. Model III does not indicate differences among the age groups, but, as Model I did for Dataset B, does show the expected effect for the linguistic factor, word length. The model does not find a statistically significant

TABLE 8 | Logistic mixed-effect regression models for (ING) in Dataset C (516 tokens).

	Model III: human coders (*N* = 516 tokens)			**Model IV: radial SVM predictions (*N* = 516 tokens)**		
	Est.	**Std. Err.**	***p***	**Est.**	**Std. Err.**	***p***
(Intercept)	5.86	1.40	< 0.0001***	4.01	1.00	0.0001***
Corpus (DCB, vs. DCA)	0.25	0.61	0.6846	−0.45	0.45	0.3110
Gender (male, vs.female)	1.17	0.55	0.0326*	1.90	0.41	0.0000***
AgeGrp (AG1, vs. AG4)	−0.10	0.82	0.9024	−2.01	0.64	0.0017**
AgeGrp (AG2, vs. AG4)	−1.38	0.86	0.1058	−1.42	0.64	0.0271*
AgeGrp (AG3, vs. AG4)	−1.16	0.78	0.1358	−1.25	0.57	0.0283*
SEC (SE2, vs. SE1)	−1.09	0.64	0.0888.	−0.36	0.46	0.4380
SEC (SE3, vs. SE1)	−3.67	0.77	< 0.0001***	−1.04	0.48	0.0301*
GramCat (N-like, vs. V-like)	−0.93	0.49	0.0573.	−0.68	0.38	0.0792.
Word Len (# Sylls)	−1.39	0.42	0.0010***	−0.66	0.30	0.0286*

****p < 0.001, **p < 0.01, *p < 0.05, .p < 0.1.*

difference for grammatical category, although the data trend in the expected direction. Model IV, for the radial SVM coded data, also identifies a significant effect for gender, with greater use of *-in* for males. The effects for socioeconomic status are less robust than for the human coded data, but still in the same direction with a significant difference between rates for the highest SEC group in comparison to the lowest. Model IV matches Model III in obtaining a non-significant trend for grammatical category. A major contrast between the two models is that in comparison to the human coded data, the SVM results overemphasize age differences, suggesting significant differences among the age groups that were not seen in Model III. Finally, the SVM-coded data identifies the significant syllable length effect, which the forced alignment model in section Coding via Forced Alignment did not appear to be sensitive to. This is an indication that, unlike the unsupervised forced alignment data, the classifiers trained on hand-coded data did learn associations related to the perceptions of human coders.

(ING) Classification, Using Variable-Adjacent Productions as Training Data

In this section we return to the idea implemented by Yuan and Liberman (2009, 2011a) and McLarty et al. (2019), where a variable classifier might be trained on related, non-variable but "variable-adjacent" phonetic material. As discussed earlier, this is a novel suggestion with much promise, although, as raised by Villarreal et al. (2020), one that needs extensive validation before we know how much we might trust automated coding procedures that are not trained on data from the same variable contexts that they are used to classify. Here we use the *IN* and *ING* data – from non-variable word final instances of [ɪn] (words like *begin* and *win*) and [ɪŋ] (monosyllabic words like *thing* and *wing*) – as training data for (ING) classifiers. The key question is whether such forms, which are outside of the variable context and thus should provide stable acoustic evidence for forms phonetically similar to the variable productions of (ING), provide data of value for the training of a variable classifier. If these forms suffice for model training we might be able to get around the costly, slow, work-intensive step of hand-coding training data in the first place. We examine this here, by training a set of classifiers on the non-variable words and then assessing the classifiers' accuracy on the hand-coded (ING) words.

As described in the section CORAAL and its (ING) Data, our extracted MFCC dataset included 8,255 non-variable *IN* words (words like *begin*, *win*, and the word *in*) and 1,436 *ING* words (e.g., *thing, bring*, and *spring*). There are reasons to expect that these non-variable words will not form perfect approximations of the pronunciation of *-in* and *-ing* variants of (ING), however their basic phonological forms are close to the realizations relevant to variable (ING). Also, and as might be expected, the words in these classes are of greatly varying frequency with e.g., 1,054 tokens of *thing* and 7,860 instances of *in*. This could be a problem for their use as training data. Preliminary testing indicated that using different subsets of the variable-adjacent forms in our training data led to major differences in performance. One area that will need further exploration is how to prune these kinds of datasets for the most appropriate training examples. For the analysis here, we randomly subsampled 1,436 tokens from the *IN* words to match the smaller set of (1,436) *ING* words. This provided a training dataset of 2,872 words, evenly balanced for non-variable *IN* and *ING* words.

We then built three classifiers, again, a linear SVM, a radial SVM, and an RF. Each classifier was trained and tuned using 10-fold cross-validation with training on the categorical *IN* and *ING* data. We then tested each of these trained models against the 803 three-analyst hand-coded (ING) instances in Dataset B and the 516 tokens in Dataset C. The outcomes from these testing runs are presented in **Table 9**. Performance for these classifiers is slightly lower than the classifiers trained on human coded data (in section Coding via Forced Alignment), especially in their testing performance on Dataset B (comparing left-hand panels of **Tables 7**, **9**). The reduction of performance on Dataset B makes sense given that the earlier models were trained and tested on speech from the same speakers. Testing on Dataset C for the models trained on the non-variable data actually shows much less over-prediction of *-in* and only very small reduction in performance compared to the classifiers trained on Dataset B. In

TABLE 9 | Performance of the three classifiers for classification of variable (ING).

	Dataset B test performance (training on non-variable *IN* and *ING* data) **Actual *-in* rates: 62.4%**						**Dataset C performance (training on non-variable *IN* and *ING* data)** **Actual *-in* rates: 60.3%**					
	Pred. *-in* rate	**Accuracy**	**Precision**	**Recall**	***F*1 score**	**AUC**	**Pred. *-in* rate**	**Accuracy**	**Precision**	**Recall**	***F*1 score**	**AUC**
Linear SVM	64.6%	69.6%	0.77	0.75	2.99	0.67	61.8%	69.8%	0.76	0.74	2.97	0.68
Radial SVM	64.6%	78.8%	0.85	0.82	3.28	0.77	60.3%	75.2%	0.79	0.79	3.18	0.74
RF	70.0%	75.5%	0.86	0.77	3.08	0.72	64.5%	76.4%	0.84	0.78	3.14	0.74

fact, some metrics, such as the F1 Scores for the radial SVM and RF actually slightly improve; in terms of overall predicted rates of *-in*, the models trained on the non-variable tokens get closest to the actual rates for Dataset C. This result likely comes about for a couple of related reasons. First, the non-variable training data included speech from across all of the speakers in the corpus. This likely helped the models in making predictions for Dataset C, which contained variable (ING) tokens from across a wide range of speakers. Dataset B, with tokens selected from only a subset of speakers, was less useful as training data in this way, even though the hand-coded data provided clearer evidence for the models about the mappings between the features (MFCCs) and variants of (ING). Thus, as we will return to, each of the approaches here appears to have advantages, as well as disadvantages.

As a final assessment of classifier performance, we once again conduct logistic regression analyses comparing the human analysts' codes to the predictions of the classifier, again focusing on the radial SVM and on Dataset C (We do not include Dataset B simply for space). **Table 10** presents the model of this radial SVM output, Model V, along with Model III, of the human coded data (from **Table 8**). The patterns emerging from the radial SVM's predictions, as indicated by Model V, are similar to those for the human coded data for SEC, gender, and the word length effect (all yielding significant effects), and for grammatical category (not significant). However, like the SVM trained on the hand-coded data, the SVM here also identifies significant age group differences that do not emerge among the human coded data. Most notably, unlike any of the other models, Model V identifies a significant difference between the CORAAL components, with DCB speakers using less *-in* than DCA speakers.

DISCUSSION AND CONCLUSION

To conclude, each of these assessments has demonstrated that automatic coding algorithms, through both forced alignment algorithms and machine learning classifiers, can perform close to human coders in their ability to categorize the sociolinguistic variable (ING). Overall, the models tend to over-predict *-in* somewhat, but their performance is promising. In addition to achieving generally reasonable accuracy, precision, and recall, we would argue that the statistical model assessments of the sociolinguistic patterns that would be uncovered through any of these datasets tell roughly similar stories, albeit with small substantive differences (e.g., the word length effect missing from the forced alignment coded data, the SVM trained on variable-adjacent data suggesting a difference between the CORAAL components).

Given the similar performance of the automated coding approaches tested in our study, and more generally the techniques used to implement them, we suggest that the most appropriate approach for automatic coding of variable features will depend on the kind of data used, whether there is any hand coding available, and, crucially, the research questions and design of the study. Using any particular machine learning approach, such as SVMs *or* RFs, should take into account what is gained or lost in that choice. For example, with data that has acoustic measures characterized by collinearity, RFs might be preferred (Villarreal et al., 2020). At the same time, forced alignment-based classification holds great promise for use cases where an entire large corpus can be force aligned through a training and alignment process (as we did for CORAAL). This approach may be less useful or appropriate if less speech is available for training the aligner. (Although here we do note that Bailey's (2016) investigation yielded good results for variables in British English data using an aligner trained on American English.)

For (ING) in particular, in terms of the machine learning approaches, both SVMs and RFs performed relatively similarly, although across the board the radial SVM performed best, followed closely by the RFs. Models generally performed well across all performance measures, although we note that model precision was uniformly better than recall. The consistent better performance of the radial SVMs over linear SVMs indicates that automated coding methods should not be used "off the shelf," without careful testing and adjustment for the problem at hand. It may be that radial SVMs will consistently outperform linear SVMs on (socio)linguistic data – this would not be surprising – but individual projects should assess that empirically.

Further, our approach to more customized classifiers worked slightly better with hand-coded training data than it did with adjacent, non-variable productions as the training data. Thus, and not surprisingly, it would seem most prudent to use validated carefully hand-coded data for model training when it is available – and we would argue that using such data for testing is crucial – but our results should be taken as additional encouraging evidence, building on Yuan and Liberman (2009, 2011a) and McLarty et al. (2019), that using adjacent, non-variable training data can hold good promise in certain cases. Of course, some variables will be more appropriate to examine through this means than others, where it may prove impossible to identify relevant non-variable analogs. Therefore, we encourage

TABLE 10 | Logistic mixed-effect regression models for (ING) in Dataset C (516 tokens) (Model III repeated from **Table 8**).

	Model III: human coders (*N* = 516 tokens)			**Model V: radial SVM predictions (*N* = 516 tokens)**		
	Est.	**Std. Err.**	***p***	**Est.**	**Std. Err.**	***p***
(Intercept)	5.86	1.40	< 0.0001***	3.47	0.75	< 0.0001***
Corpus (DCB, vs. DCA)	0.25	0.61	0.6846	−0.69	0.32	0.0334*
Gender (male, vs.female)	1.17	0.55	0.0326*	0.66	0.27	0.0155*
AgeGrp (AG1, vs. AG4)	−0.10	0.82	0.9024	−1.11	0.44	0.0126*
AgeGrp (AG2, vs. AG4)	−1.38	0.86	0.1058	−1.30	0.43	0.0026**
AgeGrp (AG3, vs. AG4)	−1.16	0.78	0.1358	−0.44	0.37	0.2362
SEC (SE2, vs. SE1)	−1.09	0.64	0.0888.	0.01	0.32	0.9704
SEC (SE3, vs. SE1)	−3.67	0.77	< 0.0001***	−0.83	0.34	0.0140*
GramCat (N-like, vs. V-like)	−0.93	0.49	0.0573.	−0.34	0.31	0.2643
Word Len (# Sylls)	−1.39	0.42	0.0010***	−0.75	0.25	0.0022**

****p* < 0.001, ***p* < 0.01, **p* < 0.05, .*p* < 0.1.*

analysts considering automating coding to carefully consider the details of their particular use case when determining which method and which type of training data are most appropriate for their situation.

One major take-away from our study is that rather than interpret automated techniques' performance as measured against some notion of perfect accuracy, we find that human coding for the variable also achieves agreement at rates of only about 88%. We need to ask what it would mean for an automated system to perform better than this. To return to questions this paper began with: on what basis should algorithms' performance be assessed? What counts as successful? And, by what metric? It would seem that accuracy as measured against a set of human coded data (especially by a single coder) is not the right metric (unless one's goal is to replicate exactly the coding practices of an individual analyst). Rather, measures of performance, and success, should recognize that "gold standard" sociolinguistic data are inherently variable, not just in the patterns in the data but also in the practices used for assessment, even in the best of cases. Our comments here parallel conclusions from work on other aspects of linguistic coding and annotation, such as phonetic transcription (e.g., Shriberg and Lof, 1991; Cucchiarini, 1996). Regardless of the specific problem, success should ultimately be measured in terms of the adequacy of the resultant data for the purposes at hand. Similar to Reddy and Stanford's (2015) discussion of desiderata for a fully automated vowel measurement system (see also Kendall and Vaughn, 2020 and Kendall and Fridland, 2021: chapter 8), we argue that automated techniques should not seek simply to replace human analysts but rather that they reflect alternative approaches to coding that have advantages and appropriateness for some applications and disadvantages and inappropriateness for others.

Across automated techniques our results largely triangulate toward a reasonable and not unexpected pattern for (ING) in CORAAL. But of course differences between the individual models' performances and the statistical patterns that emerge caution against taking an uncritical view of, for instance, a *p*-value threshold as the measure of patterns in a dataset. That said, we believe the variability in human coded data offers a similar caution. That is, our statistical analyses of human coded (ING) in CORAAL (in Models I and III) also present somewhat different views of the patterns in CORAAL. It would seem that the story they tell in the aggregate provides a more dependable picture of the patterns for this sociolinguistic variable in these data than any single one of the models. This seems to us a useful observation for the larger sociolinguistic enterprise and not just a point relevant to automatic coding procedures.

As a final note, we observe that major advances in other domains of machine learning and artificial intelligence, such as automatic speech recognition, have come about through the development of larger and larger "gold standard" training datasets. We stress the importance and value of carefully hand-coded datasets, along with the understanding of the variation inherent in auditory coding. Until automated procedures have been extensively validated for a wide range of features and datasets – something that appears to be still rather far off in the future – a bottleneck in the advancing of automated procedures will remain the availability of hand-coded training data. Our study has focused on relatively small training data, and this seems to us an important area for sociolinguistics at present, since large, reliable human coded datasets for variables like (ING) are unlikely to be available in the immediate future. However, it stands to reason that the performance of these kinds of automated classification systems will be improvable with larger training data, which we believe presents a call for greater data-sharing and organized efforts toward open science in the field. Thus, readers will find our datasets, as well as code, included in the **Supplementary Material** with this paper.

DATA AVAILABILITY STATEMENT

The datasets presented in this study come from the public Corpus of Regional African American Language, which is available at https://oraal.uoregon.edu/coraal. All derived data along with processing code are provided as **Supplementary Material** with this paper.

AUTHOR CONTRIBUTIONS

TK and CV conceived of and designed the study, with collaboration from CF, KG, JM, CT, and SA. TK, CV, KG, and CF drafted and revised the paper, with help from JM and CT. JM and CT contributed the most human data coding to the project, with TK, CF, CV, KG, and SA also contributed data coding. SA, with help by TK and JM, generated part-of-speech annotations used in the paper. CF managed the forced alignment process. TK conducted the machine learning procedures, with help from CV, KG, and CF. All authors contributed to the article and approved the submitted version.

FUNDING

This project, as well as the Corpus of Regional African American Language (CORAAL) more generally, was made possible with support from U.S. National Science Foundation grant # BCS-1358724.

ACKNOWLEDGMENTS

We thank Michael McAuliffe for helpful advice and feedback about MFA. We also thank Morgan Sonderegger and the editors of the Frontiers Computational Sociolinguistics Research Topic for helpful and motivating discussions. Finally, we thank our editor, Jussi Karlgren, and the two reviewers for helpful feedback on our manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.648543/full#supplementary-material

This includes code and data to replicate and further explore the studies presented here. See the Guide to **Supplementary Material** for more details.

REFERENCES

Arnson, S., Kendall, T., Farrington, C., and McLean, J. (in progress). *Part of Speech Tagged Version of CORAAL*. Eugene, OR: The Online Resources for African American Language Project.

Bailey, G. (2016). *Automatic detection of sociolinguistic variation using forced alignment*. University of Pennsylvania Working Papers in Linguistics, Vol. 22, Article 3. Penn Libraries, Philadelphia, PA, United States.

Binnenpoorte, D. M. (2006). *Phonetic transcriptions of large speech corpora*. Doctoral dissertation. Radboud University Nijmegen, Nijmegen, The Netherlands.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual Workshop on Computational Learning Theory* (New York, NY), 144–152. doi: 10.1145/130385.130401

Chang, C.-C., and Lin, C.-J. (2001). *LIBSVM: A library for Support Vector Machines*. National Taiwan University. Available online at: http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed December 15, 2020).

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychol. Bull*. 88, 322–328. doi: 10.1037/0033-2909.88.2.322

Cucchiarini, C. (1993). *Phonetic transcription: A methodological and empirical study*. Doctoral dissertation. Radboud University Nijmegen, Nijmegen, The Netherlands.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clin. Linguist. Phonetics* 10, 131–155. doi: 10.3109/02699209608985167

Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust*. 28, 357–366. doi: 10.1109/TASSP.1980.1163420

Duckworth, M., McDougall, K., de Jong, G., and Shockey, L. (2011). Improving the consistency of formant measurement. *Int. J. Speech Lang. Law* 18, 35–51. doi: 10.1558/ijsll.v18i1.35

Eckert, P. (2008). Variation and the indexical field. *J. Sociolinguist*. 12, 453–476. doi: 10.1111/j.1467-9841.2008.00374.x

Farrington, C. (2018). Incomplete neutralization in African American English: the case of final consonant voicing. *Lang. Var. Change* 30, 361–383. doi: 10.1017/S0954394518000145

Farrington, C. (2019). *Language Variation and the Great Migration: Regionality and African American Language*. Ph.D. dissertation. University of Oregon, Eugene, OR, United States.

Fasold, R. W. (1972). *Tense Marking in Black English: A Linguistic and Social Analysis*. Arlington, VA: Center for Applied Linguistics.

Forrest, J. (2017). The dynamic interaction between lexical and contextual frequency: a case study of (ING). *Lang. Var. Change* 29, 129–156. doi: 10.1017/S0954394517000072

Forrest, J., and Wolfram, W. (2019). The status of (ING) in African American language. *Am. Speech* 94, 72–90. doi: 10.1215/00031283-7308049

Guy, G. (1980). "Variation in the group and the individual: the case of final stop deletion," in Locating Language in Time and Space, ed W. Labov (Waltham, MA: Academic Press), 1–36.

Hall-Lew, L., and Fix, S. (2012). Perceptual coding reliability of (L)-vocalization in casual speech data. *Lingua* 122, 794–809. doi: 10.1016/j.lingua.2011.12.005

Hazen, K. (2008). (ING): a vernacular baseline for English in Appalachia. *Am. Speech* 83, 116–140. doi: 10.1215/00031283-2008-008

Hazen, K. (2011). Flying high above the social radar: coronal stop deletion in modern Appalachia. *Lang. Var. Change* 23, 105–137. doi: 10.1017/S0954394510000220

Houston, A. C. (1985). *Continuity and Change in English Morphology: The Variable (ING)*. Ph.D. dissertation. University of Pennsylvania, Philadelphia, PA, United States.

Huang, X., Acero, A., Hon, H. W., and Reddy, R. (2001). Spoken Language Processing: A Guide to Theory, *Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall PTR.

Kendall, T. (2013). *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. Basingstoke: Palgrave Macmillan. doi: 10.1057/9781137291448

Kendall, T., and Farrington, C. (2020a). *The Corpus of Regional African American Language*. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project. Available online at: http://oraal.uoregon.edu/coraal (accessed December 15, 2020).

Kendall, T., and Farrington, C. (2020b). *CORAAL User Guide*. Version 2020.05. Eugene, OR: The Online Resources for African American Language Project.

Kendall, T., Fasold, R., Farrington, C., McLarty, J., Arnson, S., and Josler, B. (2018a). *The Corpus of Regional African American Language: DCA (Washington DC 1968)*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project.

Kendall, T., and Fridland, V. (2021). *Sociophonetics*. Cambridge: Cambridge University Press. doi: 10.1017/9781316809709

Kendall, T., Quartey, M., Farrington, C., McLarty, J., Arnson, S., and Josler, B. (2018b). *The Corpus of Regional African American Language: DCB (Washington DC 2016)*. Version 2018.10.06. Eugene, OR: The Online Resources for African American Language Project.

Kendall, T., and Thomas, E. R. (2019). "Variable (ING)," in *Mexican American English*, ed E. R. Thomas (Cambridge: Cambridge University Press), 171–197. doi: 10.1017/9781316162316.007

Kendall, T., and Vaughn, C. (2020). Exploring vowel formant estimation through simulation-based techniques. *Linguist. Vanguard* 6:20180060. doi: 10.1515/lingvan-2018-0060

Kessens, J. M., Cucchiarini, C., and Strik, H. (2003). A data-driven method for modeling pronunciation variation. *Spee Commun.* 40, 517–534. doi: 10.1016/S0167-6393(02)00150-4

Kessens, J. M., Wester, M., Cucchiarini, C., and Strik, H. (1998). "The selection of pronunciation variants: comparing the performance of man and machine," in *Proceedings of the fifth International Conference on Spoken Language Processing (ICSLP'98), Vol. 6* (Sydney), 2715–2718.

Labov, W. (1963). The social motivation of a sound change. *Word* 19, 273–309. doi: 10.1080/00437956.1963.11659799

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W. (1989). The child as linguistic historian. *Lang. Var. Change* 1, 85–97. doi: 10.1017/S0954394500000120

Labov, W. (2001). *Principles of Linguistic Change, Vol 2: Social Factors*. Oxford: Blackwell-Wiley.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language* 89, 30–65. doi: 10.1353/lan.2013.0015

Labov, W., Yaeger, M., and Steiner, R. (1972). *A Quantitative Study of Sound Change in Progress*. NSF-GS-3287. Philadelphia, PA: The U.S. Regional Survey.

Landis, J. R., and Koch, G. (1977). Observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22.

Liberman, M. Y. (2019). Corpus phonetics. *Ann. Rev. Linguist.* 5, 91–107. doi: 10.1146/annurev-linguistics-011516-033830

Ligges, U., Krey, S., Mersmann, O., and Schnackenberg, S. (2018). *tuneR: Analysis of Music and Speech*. Available online at: https://CRAN.R-project.org/package=tuneR (accessed December 15, 2020).

McAuliffe, M., Scolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). *Montreal Forced Aligner*. Available online at: https://montrealcorpustools.github.io/Montreal-Forced-Aligner/ (accessed December 15, 2020).

McLarty, J., Jones, T., and Hall, C. (2019). Corpus-based sociophonetic approaches to postvocalic r-lessness in African American Language. *Am. Speech* 94, 91–109. doi: 10.1215/00031283-7362239

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., et al. (2019). *Package 'e1071'*. Available online at: https://cran.r-project.org/package=e1071 (accessed December 15, 2020).

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH: Department of Psychology, Ohio State University. Available online at: www.buckeyecorpus.osu.edu (accessed December 15, 2020).

Reddy, S., and Stanford, J. (2015). Toward completely automated vowel extraction: introducing DARLA. *Linguist. Vanguard* 1, 15–28. doi: 10.1515/lingvan-2015-0002

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., et al. (2014). *FAVE (Forced Alignment and Vowel Extraction) Program Suite* v1.2.2. Philadelphia, PA: University of Pennsylvania.

Schuppler, B., Ernestus, M., Scharenborg, O., and Boves, L. (2011). Acoustic reduction in conversational Dutch: a quantitative analysis based on automatically generated segmental transcriptions. *J. Phon.* 39, 96–109. doi: 10.1016/j.wocn.2010.11.006

Shriberg, L. D., and Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clin. Linguist. Phon.* 5, 225–279. doi: 10.3109/02699209108986113

Shuy, R., Wolfram, W., and Riley, W. K. (1968). *Field Techniques in an Urban Language Study*. Washington, DC: Center for Applied Linguistics.

Sonderegger, M., Stuart-Smith, J., McAuliffe, M., Macdonald, R., and Kendall, T. (in press). "Managing data for integrated speech corpus analysis in SPeech Across Dialects of English (SPADE)," in *Open Handbook of Linguistic Data Management*, eds A. Berez-Kroeker, B. McDonnell, E. Koller, and L. Collister (Cambridge: MIT Press).

Stuart-Smith, J. (2007). "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Laboratory Phonology 9*, eds J. Cole and J. I. Haulde (New York, NY: Mouton de Gruyter), 65–86.

Tagliamonte, S. (2004). "Someth[in]'s go[ing] on!: variable ing at ground zero," in *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe, ICLaVE 2*, eds B.-L. Gunnarsson, L. Bergström, G. Eklund, S. Fidell, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundergren, and M. Thelander, (Uppsala, Sweden: Uppsala Universitet), 390–403.

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of York English: was/were variation as a case study for statistical practice. *Lang. Var. Change* 24, 135–178. doi: 10.1017/S0954394512000129

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge: University of Cambridge Press.

Van Bael, C., Boves, L., van den Heuvel, H., and Strik, H. (2007). Automatic phonetic transcription of large speech corpora. *Comput. Speech Lang.* 21, 652–668. doi: 10.1016/j.csl.2007.03.003

Vaughn, C., and Kendall, T. (2018). Listener sensitivity to probabilistic conditioning of sociolinguistic variables: the case of (ING). *J. Mem. Lang.* 103, 58–73. doi: 10.1016/j.jml.2018.07.006

Villarreal, D., Clark, L., Hay, J., and Watson, K. (2020). From categories to gradience: auto-coding sociophonetic variation with random forests. *Lab. Phonol.* 11:6. doi: 10.5334/labphon.216

Wagner, S. E. (2012). Age grading in sociolinguistic theory. *Lang. Linguist. Compass* 6, 371–382. doi: 10.1002/lnc3.343

Weinreich, U., Labov, W., and Herzog, M. (1968). "Empirical foundations for a theory of language change," in *Directions for Historical Linguistics*, eds W. P. Lehmann and Y. Malkiel (Austin, TX: University of Texas Press), 95–195.

Wester, M., Kessens, J. M., Cucchiarini, C., and Strik, H. (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Lang. Speech* 44, 377–403. doi: 10.1177/00238309010440030401

Wolfram, W. (1993). "Identifying and interpreting variables," in *American Dialect Research*, ed D. Preston (Philadelphia/Amsterdam: Jon Benjamins), 193–221. doi: 10.1075/z.68.10wol

Wolfram, W. A. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

Yaeger-Dror, M., Kendall, T., Foulkes, P., Watt, D., Oddie, J., Harrison, P., et al. (2009). "Perception of r-fulness by trained listeners," in *Paper presented at the Linguistic Society of America* (San Francisco, CA).

Yuan, J., and Liberman, M. Y. (2008). Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* 123:3878. doi: 10.1121/1.2935783

Yuan, J., and Liberman, M. Y. (2009). Investigating /l/ variation in English through forced alignment. *Proc. Interspeech* 2009, 2215–2218.

Yuan, J., and Liberman, M. Y. (2011a). /l/ variation in American English: a corpus approach. *J. Speech Sci.* 1, 35–46. doi: 10.20396/joss.v1i2.15025

Yuan, J., and Liberman, M. Y. (2011b). "Automatic detection of "g-dropping" in American English using forced alignment," in *2011 Proceedings of Automatic Speech Recognition and Understanding* (Waikoloa, HI), 490–493. doi: 10.1109/ASRU.2011.6163980

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/frai.2021.642505

Reduction of Survey Sites in Dialectology: A New Methodology Based on Clustering

Péter Jeszenszky, Carina Steiner and Adrian Leemann*

Center for the Study of Language and Society, Faculty of Humanities, University of Bern, Bern, Switzerland

Many language change studies aim for a partial revisitation, i.e., selecting survey sites from previous dialect studies. The central issue of survey site reduction, however, has often been addressed only qualitatively. Cluster analysis offers an innovative means of identifying the most representative survey sites among a set of original survey sites. In this paper, we present a general methodology for finding representative sites for an intended study, potentially applicable to any collection of data about dialects or linguistic variation. We elaborate the quantitative steps of the proposed methodology in the context of the "Linguistic Atlas of Japan" (LAJ). Next, we demonstrate the full application of the methodology on the "Linguistic Atlas of German-speaking Switzerland" (*Germ.:* "Sprachatlas der Deutschen Schweiz"—SDS), with the explicit aim of selecting survey sites corresponding to the aims of the current project "Swiss German Dialects Across Time and Space" (SDATS), which revisits SDS 70 years later. We find that depending on the circumstances and requirements of a study, the proposed methodology, introducing cluster analysis into the survey site reduction process, allows for a greater objectivity in comparison to traditional approaches. We suggest, however, that the suitability of any set of candidate survey sites resulting from the proposed methodology be rigorously revised by experts due to potential incongruences, such as the overlap of objectives and variables across the original and intended studies and ongoing dialect change.

Keywords: dialectology, survey site selection, subsampling, clustering, language variation and change, dialect survey, linguistic geography

Edited by:
Dirk Hovy,
Bocconi University, Italy

Reviewed by:
Yves Scherrer,
University of Helsinki, Finland
John Nerbonne,
University of Freiburg, Germany

****Correspondence:***
Péter Jeszenszky
peter.jeszenszky@csls.unibe.ch

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *16 December 2020*
Accepted: *24 March 2021*
Published: *20 May 2021*

Citation:
Jeszenszky P, Steiner C and Leemann A (2021) Reduction of Survey Sites in Dialectology: A New Methodology Based on Clustering. Front. Artif. Intell. 4:642505. doi: 10.3389/frai.2021.642505

1. INTRODUCTION

1.1. Motivation

Spatial sampling for a dialect study, i.e., choosing localities to survey, has been one of the central issues in dialectology. Similar to the selection of speakers, the selection of surveyed localities (termed *"survey sites"* in this paper) needs careful planning according to study criteria, such as comparability and representativeness of areas, social groups, and linguistic levels. The issue of survey site reduction is as old as surveying itself. Linguistic and, specifically, dialect studies often select their survey sites from earlier data collections, such as linguistic atlases, and choose sites where certain linguistic variables of interest have already been documented. However, site reduction is usually done qualitatively, based on linguistic expertise, without quantitative arguments supporting selection procedures. Thus, despite their importance, most methodologies used for the reduction of survey sites in dialect studies are not reproducible in detail.

Researchers can adopt more objective procedures and potentially optimize their resources by utilizing quantitative methods for locating representative survey sites. Cluster analysis techniques are especially appropriate for this task and are already known in linguistics. This paper outlines a general methodology for the problem of quantitative survey site selection based on previously recorded dialect data (e.g., linguistic atlases) and proposes an application of cluster analysis. To demonstrate the versatility of the approach, the general methodology is also applied to real dialect data sets, in one case with the aim of finding suitable survey sites for an actual contemporary dialect study.

Researchers from different areas of linguistics could potentially benefit from the methodology proposed in this paper, by utilizing survey site networks of previous studies. A potential research aim may be to conduct a dialect interview campaign, revisiting numerous phenomena in a dialect atlas, or recording new phenomena at the same linguistic level as the atlas, with the objective of covering the expected variation in fewer locations. As another example, researchers may want to test how the dialects in a certain area have changed, and so they plan to revisit a previous survey. Assuming that dialect leveling has occurred since the survey, they may only want to visit a sufficient number of sites representing the contemporary dialectal landscape. Additionally, the methodology could be implemented for larger databases based on online crawling or digitized corpora (e.g., Anderwald and Wagner, 2007; Huang et al., 2016; Ueberwasser and Stark, 2017; Grieve et al., 2019; Willis, 2020), where the researcher might need to select limited, representative survey sites after appropriate data pooling (e.g., spatially).

1.2. Research Objectives

Linguistic studies often start with the task of survey site selection based on the sites of a previous larger scale survey. Our aim is to provide general suggestions about optimal survey site subsampling to the linguistic/dialectology community. As summarized in the "first law of geography" (Tobler, 1970), variation is assumed to be spatially autocorrelated. Representing variation in linguistic space is therefore deemed to represent the variation within the underlying data in geographic space as well [cf. *Fundamental Dialectological Postulate* (FDP)—Nerbonne and Kleiweg, 2007]. Consequently, subsampling survey sites based on a spatial grid, as often done in dialectology, could theoretically represent linguistic variation. We hypothesize, however, that cluster analysis—already extensively used in dialectometry for finding representative areas and boundaries—can also be utilized for finding representative survey sites for a related or follow-up study.

We address this hypothesis based on two research objectives. First, we propose a general methodology, outlining the steps for finding suitable association measures, subsequent clustering methods and their possible validation, and, finally, a qualitative evaluation of the reduced set of survey sites. Second, we present the practical application of the methodology on the example of the "*Sprachatlas der Deutschen Schweiz*" (SDS—Hotzenköcherle et al., 1962–2003). The specific aim of this application is to reduce the number of survey sites to a representative subset of a predetermined size, to be used for a subsequent study, "Swiss German Dialects Across Time and Space" (SDATS[1]—Leemann et al., 2020c). However, dialect change and socioeconomic processes have occurred since the collection of SDS data (around 1939–1958). This application example includes appointing a candidate survey site subset resulting from the quantitative steps and qualitative revision to estimate *contemporary* dialectal variation, in correspondence to the needs of SDATS. Thus, we address a research requirement beyond finding a representative survey site set in a collection by inferring a future state of language. We argue that most studies aiming to perform survey site reduction have similar objectives and, therefore, would benefit from incorporating these considerations into their methodologies. Additionally, integrated into the outline of the general methodology, we provide a proof of concept based on the "Linguistic Atlas of Japan" (LAJ—NLRI, 1966–1974), demonstrating the breadth of its applicability.

2. BACKGROUND

2.1. Site Selection in Spatial Sciences and Dialectology

Finding point-like sampling locations (survey sites) for representing reality is a key issue in spatial sciences, and representativeness is heavily dependent on the spatial structure of the variable of interest. Effective spatial sampling has to consider the spatial autocorrelation in the population, and the variables investigated (e.g., Griffith, 2005; Kumar et al., 2011). Most linguistics surveys focus on multiple variables, necessitating a balanced sampling strategy to capture factors, such as linguistic levels, regional variation of language, and extra-linguistic factors. Practical considerations, such as available respondents and research budgets, impose further constraints on study planning. Linguistic surveys (including large-scale dialect atlases, and projects sampling their sites of interest from previous data sets) often detail their speaker selection criteria (e.g., Linn, 1983) but disclose less about selection process of their survey sites (for exceptions, see MacAulay's review, 2018).

Spatial sciences use numerous sampling strategies (cf. e.g., Ripley, 1981; Olea, 1984; Delmelle, 2009) that are already present in linguistic research. In a *random sampling* approach, each point in a population (or area) has an equal probability of being selected. At the same time, the spatial distribution patterns of linguistic phenomena do not always follow the spatial distribution of other population traits. Therefore, random sampling might lead to oversampling the variable of interest in densely populated regions where few variants prevail, or to undersampling in areas with low, isolated populations that use diverse variants. In linguistics, randomly selecting people has been, however, successfully utilized for sociolinguistic studies, as a large enough sample may be representative of the entire population (Bailey and Dyer, 1992).

Systematic or *stratified sampling* divides the population into groups (e.g., Kondo et al., 2014), often by grids in space. Sample sites within this grid (which can be square,

[1] www.sdats.ch

hexagonal, adjusted to the population, e.g., by a Voronoi-tessellation, etc.) are chosen systematically or at random. If applied spatially, stratified sampling essentially maximizes the distance between survey sites and gives less chance for undersampling, but might also oversample densely populated areas, where variation may be lower. At the same time, sparsely settled areas may also be oversampled, especially if the variation is lower, for instance, in relatively newly settled or expansion areas of a language (e.g., Western United States, Lapland, Hokkaido, Siberia). *Adjusted sampling* specifically concentrates on avoiding over- and undersampling by densifying the survey site network in areas with higher expected variation (cf. Cressie, 2015). Most traditional large-scale linguistic atlases selected their survey sites based on such spatial grids (cf. McDavid, 1971), e.g., the Slavic Linguistic Atlas ("Obščeslavjanskij lingvističeskij atlas" OLA —Avanesov, 1965), with some regional atlas projects on German dialects using coordinated survey grids ("Sprachatlas von Bayerisch-Schwaben," SBS—König, 1996–2009; "Vorarlberger Sprachatlas," VALTS—Gabriel, 1985; "Südwestdeutsche Sprachatlas," SSA—Steger and Schupp, 1993). A grid method was used for selecting the most central sites of REDE's "Digitaler Wenkeratlas" (DIWA—Lameli et al., 2015) from the original points of the Wenker Atlas. Projects using adjusted sampling include the "Sprach- und Sachatlas Italiens und der Südschweiz" (AIS—Jaberg and Jud, 1928–40), the "Linguistic Atlas of the Middle and Southern Atlantic States" (LAMSAS—Kurath, 1949; McDavid, 1971), and the "New Linguistic Atlas of Japan" (NLJ— Onishi, 2016). SDS is also a relevant example, as sampling was scaled according to linguistic variation over population density.

2.2. Grouping and Survey Site Reduction in Dialectology

In a site reduction task, a reduced number of sites are selected from existing samples, such that they are representative of other sites, typically in their neighborhood (cf. Olea, 1984). Computational science provides an extensive coverage of problems related to selecting data points that efficiently describe an entire data set (e.g., Daszykowski et al., 2002; Elhamifar et al., 2012; Gani and Limam, 2016). Spatial sciences (such as soil science and vegetation ecology) and fields where the distribution and change of variables over time are also spatially autocorrelated provide various site reduction methods. For example, Lengyel et al. (2011) select subsets of their vegetation plots by sorting them based on decreasing mean dissimilarity between pairs and then sorted again by increasing variance of these dissimilarities. While many site reduction methods in the spatial sciences focus on finding a subsample for optimizing the extraction of one or a few variables (such as soil attributes, e.g., Maltauro et al., 2019, or species abundance, e.g., Loos et al., 2015), linguistic studies might aim to be representative of tens or hundreds of linguistic variables. Besides, proximity in space *per se* does not define dialect similarity (cf. Szmrecsanyi, 2012), and people, the agents of linguistic variation, are constantly on the move, contributing to a changing spatial distribution of linguistic variables.

Linguistic studies aiming at the comparison of contemporary and older data, however, need to revisit all or a reduced subset of the original survey sites. It is intuitive to convey patterns and trends by grouping sites together for example, by drawing isoglosses and naming dialect areas. According to the law of spatial autocorrelation, nearer sites are expected to be similar and distant ones to be dissimilar (Tobler, 1970; Legendre, 1993; Nerbonne and Kleiweg, 2007). This general correlation is often confirmed in dialectology. Cluster analysis, the quantitative grouping of data, resulting in a lower number of representative groups, is also a fundamental procedure in dialectometry. The general procedure of data analysis in standard modern dialectometry involves the calculation of linguistic distances between every pair of survey sites, producing a linguistic distance matrix. This matrix is then analyzed using a variety of multivariate statistics, including multidimensional scaling and cluster analysis, to identify common patterns of regional variation (Grieve, 2014).

Site reduction can be considered a similar task to finding groups and patterns among survey sites. Most projects that apply site reduction to select sites from earlier collections, usually select their sites, such that they retain the spatial density of sites in the original study (e.g., Séguy, 1973; Kelle, 2001; Bucheli and Glaser, 2002; Lameli et al., 2015; Onishi, 2016; Budin et al., 2019). Spatial autocorrelation is usually assumed without quantitative testing, and the sites are verified case-by-case, introducing potential subjectivity and untested representation. Despite the availability of sophisticated methods for deriving dialect areas and spatial patterns, these methods have not often been used for site reduction.

The methodology presented in this paper fills this research gap, by demonstrating the value of cluster analysis for the task of survey site reduction from previous collections of data.

2.2.1. Cluster Analysis

Most clustering procedures take association matrices (such as linguistic distance matrices) as inputs, based on which clusters are compared (Borcard et al., 2011). There are two relevant clustering techniques important for the methodology in this paper, distinguished by the underlying clustering algorithms, necessary inputs, and analytical procedures. *Hierarchical clustering* is the family of clustering methods mainly used in dialectometry. Its algorithms build a hierarchy among the data points in a nested sequence of partitions (see overviews in Heeringa 2004, p. 146–156; Nerbonne et al. 2008; Levshina 2015, p. 309–311). In hierarchical clustering, every step splits an existing cluster in two, based on a certain metric. Importantly, a linkage criterion is needed to specify the dissimilarity between the clusters present and the newly formed cluster. *Partitional clustering*, usually not used in dialectometry (Nerbonne and Wieling, 2018), aims at breaking the data set into a predetermined number of groups and finds these groups simultaneously, refining the solution in every iteration. Although partitional algorithms disregard hierarchy within the classification, Prokić and Nerbonne (2008) find that the results of the *k*-means partitioning algorithm correspond to dialectal divisions made by experts.

We introduce three clustering methods that are generally considered to perform well in dialectology. According to the arguments of several scholars in dialectology (Heeringa, 2004; Prokić and Nerbonne, 2008; Grieve et al., 2011; Syrjänen et al., 2016; Burridge et al., 2019; Lameli et al., 2020), we decided to test the two most promising hierarchical clustering algorithms and one partitional algorithm. Algorithms in hierarchical clustering differ with regard to their linkage criteria (reviewed in Jain and Dubes, 1988). The *Unweighted Pair Group Method using Arithmetic averages* (UPGMA) method assesses the dissimilarity between the new cluster and the existing cluster based on the distance between their means. In this process, each element in a cluster gets an equal weight, independent of the number of elements in the clusters (Sneath and Sokal, 1973, p. 228). *Ward's algorithm* (1963) works differently with regard to the linkage criterion. It minimizes the within-cluster variance and therefore is prone to producing compact clusters of similar size (within the dimension of linguistic distances) (Wilks, 1995), which is not always reasonable. Grieve et al. (2011) use Ward's method because it is based on the analysis of variance, while Prokić and Nerbonne (2008) find the UPGMA and Ward's method to perform best for dialectometry. Heeringa (2004) provides a comparison between the UPGMA and Ward's method, but finds UPGMA to perform better on Dutch dialect data.

The third algorithm selected is the *Partitioning Around Medoids* (PAM) algorithm (Kaufman and Rousseeuw, 1987), a popular algorithm for clustering non-Euclidean data (Schubert and Rousseeuw, 2019). As a partitioning clustering method, PAM classifies all observations within a data set into k number of clusters specified beforehand. The main difference between PAM (also known as k-medoids algorithm) and the widely used k-means algorithm is that in each step, PAM appoints actual data points (medoids) as the centers of clusters by minimizing the distance between the points and the medoid. K-means, however, minimizes the sum of squared Euclidean distances, which makes it less robust to noise and outliers than k-medoids (Park and Jun, 2009). Partitioning algorithms are not commonly used in dialectology. However, k-means was applied by Hyvönen et al. (2007) and Burridge et al. (2019), while Cheshire et al. (2011) and Syrjänen et al. (2016) applied k-medoid clustering on different kinds of linguistic data.

A general problem of clustering procedures is that they always deliver clusters, even if the underlying data has little clustering tendency (e.g., due to dialect continua). Hierarchical clustering is, additionally, prone to large differences in results caused by small changes in the input matrix (cf. Jain and Dubes, 1988; Nerbonne et al., 2008). Therefore, in all cases, clustering procedures need validation in order to obtain stable and interpretable clustering results. Phylogenetic literature (Felsenstein, 2004) and dialectometry (Mucha and Haimerl, 2005; Manni et al., 2006) recommend *bootstrapping*. In dialectometry, bootstrapping resamples a data set with replacement and runs the clustering algorithm for each resampled set, arriving at a "composite" result with information about its stability (Nerbonne et al., 2008). Another popular method, *noisy clustering* builds validation in the clustering procedure by adding noise to the data to test its impact. The advantage of noisy clustering over bootstrapping is that it is also applicable to single distance matrices (Prokić and Nerbonne, 2008). The *cophenetic correlation coefficient* (Sokal and Rohlf, 1962) is often used to measure the correlation between the distances in the original data and the distances as implied by hierarchical clustering results (Heeringa, 2004; Birkenes, 2019). Further internal measures for cluster validation assess the compactness, connectedness, and separation of partitions, including the Dunn-index, which identifies compact (small variance between members) and well-separated clusters (Dunn, 1974).

External evaluation of clustering methods is often undertaken in dialectology through comparing cluster solutions to a gold standard (Heeringa et al., 2002; Prokić and Nerbonne, 2008; Lameli et al., 2020), e.g., to a meticulous qualitative dialect division made by experts. Prokić and Nerbonne (2008) compare the clustering solutions of several algorithms to a benchmark of Bulgarian dialects using the Rand-index (Rand, 1971), entropy, and purity of clusters. This kind of external evaluation may not be available for many potential studies, as the intended number of clusters might not match expert classifications of dialect areas. Meilă's variation of information (*VI*) metric (Meila, 2007), related to the entropy in clusters, compares the similarity of any two clustering partitions, approximating the human intuition of distance.

2.3. The Project SDATS

In this paper, we apply the suggested clustering-based site reduction approach suggested to the monumental SDS. This application specifically intends to consider the sociolinguistic aims and other requirements of our project SDATS (Leemann et al., 2020c).

"Swiss German Dialects Across Time and Space" aims at conducting a large-scale collection of the contemporary dialects of Swiss German and a subsequent comparison to dialectal forms recorded in SDS. To reach these goals, SDATS maintains a similar number of participants as SDS (1,000 participants, compared to c. 1,500 in SDS)[2] recruited from a reduced number of survey sites. Instead of 573 sites in the SDS, SDATS includes 125 survey sites and increases the number of speakers per site to eight speakers (of different social backgrounds) from the 1–3 "Non-mobile Old Rural Males" (and females) recorded in SDS. The main reasons for the site reduction are trends of dialect change in the last 70 years (significant leveling occurred—cf. Christen, 1998), sociolinguistic aims, manpower, and financial resources. Rather than searching for the "base dialect," as SDS did, SDATS aims to record more intralocal, colloquial variation by interviewing respondents of different backgrounds, with an emphasis on the provenance of respondents. Data collection began in 2020 by means of a custom-developed open-source smartphone application (Leemann et al., 2020b), used mostly in virtual settings (Leemann et al., 2020a) due to the COVID-19 pandemic.

Previous site reduction attempts on SDS have been arbitrary and not replicable. Kelle (2001) digitized 170 SDS maps and selected about one-sixth (101) of the original 573 survey sites, as equidistant as possible, in order to perform a new

[2] Trüb (2003) mentions that SDS had around 1,500 participants in total, but not the whole questionnaire was answered by all of them. Often the local variety is summarized based on the answers of 2–3 participants answering different parts of the questionnaire (https://sprachatlas.ch/originalmaterial-split/infos).

typological classification and confront traditional qualitative dialect classifications. Unfortunately, his selection criteria are not elaborated (with the exception of equidistance), and site representativeness could not be evaluated without digital data of the whole corpus. Almost in parallel, the "Syntactic Atlas of Swiss German" (SADS—Bucheli and Glaser, 2002) reduced the network of SDS to 383 survey sites. Their selection aimed to keep the comparison of desired isoglosses possible (Glaser and Bart, 2015), and was mainly based on merging villages with a smaller number of inhabitants into single survey sites (Bucheli Berger, 2008).

2.3.1. Digitized SDS Data

Despite being the most comprehensive collection of Swiss German dialect data, SDS has not yet been entirely digitized. Starting in 2007, Yves Scherrer (with the help of his colleagues) undertook the partial digitization of SDS for the sake of several projects; (Scherrer, 2021, 2012; Kellerhals, 2014; Scherrer and Stoeckle, 2016). In Scherrer's process, a subset of variables was defined according to linguistic criteria, with general preference given to phonological and morphosyntactic phenomena. In addition, lexical phenomena that were expected to occur frequently, such as function words, were included (Scherrer, 2021). After scanning and georeferencing the SDS maps, they appointed the locally recorded linguistic variant(s) for each survey site in each map, using geographic information systems. This procedure registered the presence and absence of each variant in digital tables. Scherrer's projects involved a simplified categorization of variables (Scherrer, 2021). This categorization granularity is in many cases (including phonetic variables), not sufficient for SDATS, which aims at a fine-grained comparison across SDS and contemporary dialect usage.

3. GENERAL METHODOLOGY

This section details a general methodology that researchers may consider for a survey site reduction task in order to identify representative sites based on data from a previous, larger-scale dialect study. At each step of the methodology, requirements and possible methods are described, and typical quantitative steps are demonstrated on the example of the LAJ (NLRI, 1966–1974). Then, in section 4, the methodology is applied to SDS data, with the specific goal of appointing survey sites for SDATS.

3.1. Requirements and the Steps of the Reduction Process

The general survey site reduction process combines the following quantitative and qualitative steps:

1. Digitize the original database, prepare the linguistic data for the sampling, typically including the (re-)categorization of variants, and select linguistic items appropriate to represent the original data in consideration of the intended study (this step is not explained in detail)
2. Calculate linguistic distance matrices based on the selected linguistic items, thus obtaining association measures among the survey sites, as detailed in section 3.2
3. Carry out the clustering procedures and appoint candidate survey sites in the resulting clusters. Typically, this step involves clustering survey sites based on one or multiple linguistic distance matrices and performing validation tests on clusters. The reduction and the subsequent selection of candidate sites are detailed in sections 3.3 and 3.4
4. Evaluate the candidate survey sites, involving (typically qualitative) revision by dialect experts and through sociogeographic filtering to find sites that correspond to the criteria of the intended study, detailed in section 3.5

To aid researchers potentially implementing this outline in their flow of research, we add a non-exhaustive list of further considerations. Our methodology assumes that the original study is part of a large-scale dialect survey. The correspondence of overlapping items in the intended study and the original data needs to be scrutinized and potentially recategorized. As it appears to be a typical task to infer contemporary dialectal variation from the original data, it is important to select items that are representative at both points in time. Thus, items that are irrelevant for the intended study should be removed, such as names of rural work-tools in a large-scale study of vernaculars. The effect of each variable or groups of variables can be tested by, e.g., jackknifing or other cross-validation methods. The selection of data will always depend on the research question, thus in some cases one or a combination of linguistic levels will be used. Although it is crucial from the point of view of data quality, we do not detail the steps of digitization in this general methodology and we assume that the original data is already digitally available.

The core of the site reduction methodology is a grouping algorithm, which classifies the survey sites within the original database into (a desired amount of) groups, with the aim of finding candidate survey sites in the resulting groups, similar to stratified sampling strategies. As in geospatial analysis, no single resampling strategy is optimal or superior: the method for subsampling also has to be appropriately selected depending on the objectives of the intended study and the original data (cf. Knollová et al., 2005). It is crucial for a researcher to decide what they mean by representativeness when selecting candidate survey sites, e.g., linguistic centrality, spatial centrality, or other, external characteristics. These decisions can be prompted by conducting exploratory analyses on the digitally available data, for example, based on aggregate linguistic distance matrices, visualizing overlaps, or testing clustering tendencies (Lawson and Jurs, 1990).

If the intended study aims to compare findings over time, then selected survey sites should already be present in the original database. Further, beyond the scope of the original data set, the selected survey sites should be representative of the survey sites surrounding them (in a linguistic sense) *at the time of the intended study*. A crucial consideration about the preservation of variation is that site reduction will always eliminate some source of variation, especially with language change occurring since the recording of the original survey. If the goal is capturing diversity, or documenting all linguistic variation possible at the expense of overall representativeness, then field knowledge and qualitative revision are crucial, as even original data or digitized data might not cover all variation. Although the clustering procedure should produce results representative of the original survey, the qualitative evaluation step might overwrite these choices.

3.1.1. The Linguistic Atlas of Japan

Typical quantitative steps in the proposed methodology are demonstrated using data from the LAJ (NLRI, 1966–1974), the largest systematic nationwide dialect collection in Japan. LAJ presents the recorded material of a large-scale survey conducted between 1957 and 1965. In total, 2,400 localities were surveyed across Japan, interviewing one (generally) male speaker per locality, born between 1879 and 1903. The atlas survey contains 285, mostly lexical phenomena.

The data set used in this example contains 37 publicly available[3] lexical variables from LAJ (Kumagai, 2016), with a focus on basic vocabulary in relation to body parts, weather and time, animals and plants, and levels of kinship. Admittedly, the focus of the data is a risk factor to results being representative of the complete lexical level recorded in LAJ.

To prepare the survey sites for a representative clustering, well-known outliers are removed, leaving 2,238 survey sites. The Ryukyu Islands, in the southwest, are removed due to their large linguistic distance from other parts of Japan. Hokkaido, in the north, was settled by the Japanese primarily as of the end of the 19th century, and thus is removed due to small dialectal variation and mixture.

3.2. Linguistic Distance

There should be significant overlap between the original data and the intended study. If that cannot be achieved, a distribution of linguistic data balanced across linguistic levels might be beneficial. Similar to clustering in dialectometry, it is advisable to take as many variables as possible from the original data set, curated for the objectives of the intended study and categorized accordingly.

Once the linguistic basis of the site reduction has been determined, researchers must construct an association measure among the survey sites. In a typical case, a linguistic distance matrix is calculated in a site × site manner, based on a set of linguistic variables.

Methods of linguistic distance calculation vary depending on the linguistic level, the variants' categorization granularity, and, if involved, the details of transcription. For calculating phonetic similarity across variants, edit distances are used most often (cf. Wieling and Nerbonne, 2015). For categorical data, linguistic distance is mostly measured based on presence and absence of variants, e.g., the Hamming distance (Spruit, 2006) or Goebl's (1983) *Relative Identity Value*, calculated on pairwise matches and mismatches. At this point, it would also be possible to test the effect of single variables. Researchers may consider removing variables with spatially similar or correlating distributions as duplicates.

Aggregate linguistic distance matrices can be explored in various ways in order to explore patterns in dialectal variation and to detect outliers and potentially problematic regions. Popular methods include similarity trees, e.g., *Neighbor-net* (cf. Cysouw, 2007), multidimensional scaling (MDS) (Heeringa, 2004; Lameli et al., 2020) (both of which are included in the dialectometry support software Gabmap—Leinonen et al., 2016), or thematic maps. The latter may focus on one certain survey site or present the aggregate picture in linguistic distance maps (Goebl, 1982; Scherrer and Stoeckle, 2016). Such plots and maps, in essence, help discover clustering tendencies and gradual transitions among dialect areas (based on the limited data).

3.2.1. Linguistic Distance Calculation Applied to LAJ

For LAJ, the linguistic distance matrix is calculated using a formula based on Goebl's *Relative Identity Value* (RIV_{jk}) (1983), similar to Scherrer and Stoeckle (2016) as applied in Jeszenszky et al. (2019). For each lexical variable, the variants (up to hundreds in some cases) are categorized on two levels. First, variant categories are constructed based on phonetic similarity. Within variant categories, further distinction is made between individual variants: variants within a variant category receive a flat difference rate[4].

We use an MDS approach to discover latent clusters and dialect continua in the data. We plot the first two or three dimensions of the multidimensional scaling results and associate the first three dimensions to RGB colors and map them[5]. These visualizations show that continua are present in this data set, thus clusters with lower stability and more (spatial) overlap are expected.

3.3. Clustering

The linguistic distance matrix is the input of clustering algorithms used for site reduction. Dialectometry often uses cluster analysis to find the internally most homogeneous and externally most heterogeneous groups in dialect data. Importantly, however, clustering techniques have mostly been used to find the optimal split[6] and spatial distribution in the data, thereby often defining dialect areas. In a typical site reduction study, however, the researcher would aim for much more than the optimal number of clusters in the data.

Hierarchical clustering results in dendrograms and association values between survey sites. Dendrograms cut at the desired or optimized level can also be spatially represented by a cluster map. Partitional clustering produces a predetermined number of clusters, the optimal number of which can be determined by optimization. Researchers might not know the exact number of survey sites they want to extract from the original set, which might influence the choice of clustering method. In any case, it is worth experimenting with different numbers of clusters, also around a previously decided number,

[3] Available online at the Linguistic Atlas of Japan DataBase (LAJDB)—www.lajdb.org.

[4] For more details on the database and linguistic distance calculation, see Kumagai (2016) and Jeszenszky et al. (2019).

[5] For a visualization of the dialects and linguistic distance in this dataset, including an MDS map, in Japan (without Okinawa), see Jeszenszky et al. (2019, p. 16–18).

[6] By means of e.g., the *silhouette technique* (Rousseeuw, 1987), it is possible to find the optimal number of clusters for partitional clustering, but it is not relevant for most studies in which the present methodology is potentially applicable, as they aim for sampling a higher number of survey sites.

in particular for exploratory analyses. Depending on their aims, researchers might determine k clusters directly for a partitional method, or, for a hierarchical method, they might select a cophenetic distance, beneath which they find their k clusters in the dendrogram. It is always possible to adjust the final number of survey sites in a qualitative revision.

3.3.1. Application of Clustering to LAJ Data

We demonstrate the performance of three clustering algorithms (PAM, UPGMA, and Ward's method), using the example of LAJ. All three clustering algorithms are implemented using the `fpc` package (Hennig, 2020)[7] in `R` (R Core Team, 2020). We perform clustering on the linguistic distance matrix resulting from section 3.2.1. Using each clustering method, partitions of $k = 20$, 50, 100, 150, 200, 300, 400, and 500 clusters are produced.

To validate the results of the different clustering methods, we used a bootstrapping approach (e.g., Nerbonne et al., 2008), as included in the `fpc` package. In the bootstrapping approach, each cluster is calculated in 100 bootstraps (default value) with resampled data. For each cluster, the Jaccard-similarities of the initial cluster solution (in bootstrap nr. 1) to cluster solutions in all other bootstraps are computed (Hennig, 2007). This approach provides *stability* values for each "composite" cluster found, based on which the performance of clustering algorithms and the sensibility of the choice of k (number of clusters) can be assessed.

As linguistic variation is assumed to be spatially autocorrelated, members of clusters found in the data are supposed to be clustered in space as well. To confirm this, and to visually explore the spatial patterns, we map the clusters produced, along with their stability values. **Figure 1** presents the clusters found in the database by the three clustering algorithms (Maps A—PAM, B—UPGMA, and C—Ward's method) with $k = 150$, an overall large number for site reduction requirements given the number of sites in LAJ. Clusters are presented on a diverging, repeating color scale. In Maps D–F, the cluster stability values are mapped to the members of the clusters. Such stability values should not be evaluated solely based on descriptive statistics (e.g., means and standard deviation) as they may vary substantially across clusters, further justifying mapping. Maps D–F also contain the histograms of the stability values, presenting considerable deviations from a normal distribution.

In PAM's map (A), clusters do not appear spatially compact, although members are clustered in the same region. In UPGMA's map (B), several clusters are visible with a high number of members. UPGMA is based on the average difference between clusters and, because of this, chaining effects are not typical for this algorithm (Lameli et al., 2020). The small, unstable clusters of single members (*singletons*) are thus possibly outliers in the linguistic space, found as clusters by this method. UPGMA shows 39 singletons for $k = 150$ while the other methods show none. Ward's method, based on their positions in the map (C) seems to find clusters structurally more similar to UPGMA. Compact clusters are a characteristic of Ward's method, contributing to its popularity in dialectology. The clusters in Map F look somewhat more stable and spatially more compact than those in PAM, and clusters of the same color (Map C) present clearer boundaries in space, appearing to overlap less. Based on the stability histograms in D–F, PAM seems to have the lowest overall values, making it less suitable for clustering than the hierarchical methods (in case of $k = 150$). It is intuitive to expect lower stability when k is higher, as smaller clusters are expected to be also more similar to neighboring clusters. This is even more significant when (dialectal) continua are present in the data, as in the case of Japanese. When applying multiple clustering algorithms with different k, it is interesting to see which algorithm produces more stable results with a certain number of clusters.

In terms of external validation,[8] Meilă's *VI* (Meila, 2007) can also be used for calculating similarities across clustering solutions. In **Table 1**, we compare the clustering solutions resulting from the three clustering methods with different k number of clusters. Higher values of Meilă's *VI* indicate greater variation between cluster solutions. Larger k usually means larger potential difference, but in our case k grows to a degree where difference between cluster solutions cannot increase anymore. Indeed, above $k = 150$, *VI* values start to decrease again. **Table 1** shows, somewhat surprisingly, that PAM is more different from the two other solutions than they are from each other, despite UPGMA's tendency for producing large clusters and singletons. This might mean that UPGMA's clusters contain, or are structurally similar to Ward's clusters, while PAM's clusters might not overlap well with Ward's. PAM's lower stability and UPGMA's and Ward's method's different cluster solutions, despite their structural similarity, suggests that researchers should strongly consider the choice of cluster algorithms.

3.4. Selecting Candidate Survey Sites

Once the validity and stability of clusters are assessed, representative sites can be identified. This can be regarded as an analogy to stratified sampling strategy, where one point is selected from each stratum. In our case, strata are the clusters, the partitions in the abstract linguistic dimensions. Studies might differ in terms of requirements for representative sites, generating several methodological considerations.

Studies might differ in the distribution granularity of variants. Resampling has to consider this granularity, along with other spatial patterns. If capturing fine-grained spatial variation is the aim, then sampling density should be adjusted accordingly. One approach could be choosing points in clusters that are central in a linguistic sense. This approach is demonstrated on LAJ in section 3.4.1 and on SDS in section 4.2.2. In case of running multiple clustering procedures and composite dendrograms, the representative sites' identity becomes less obvious, as clusters from different runs overlap. It is possible, however, to appoint a central site for each

[7] In `fpc`, the k-medoids clustering is implemented by two algorithms, `pam` is slower but produces more stable results, while `clara` is faster but more unstable. For a large number of clusters, `clara` proved significantly more unstable.

[8] Recently, the LAJ actually received a follow-up at a subset of 554 survey sites (Fukushima, 2016), in the "New Linguistic Atlas of Japan" (NLJ) (Onishi, 2016). Due to the low number of (only lexical) variables in our data set, however, it would not be sensible to use NLJ as a ground truth.

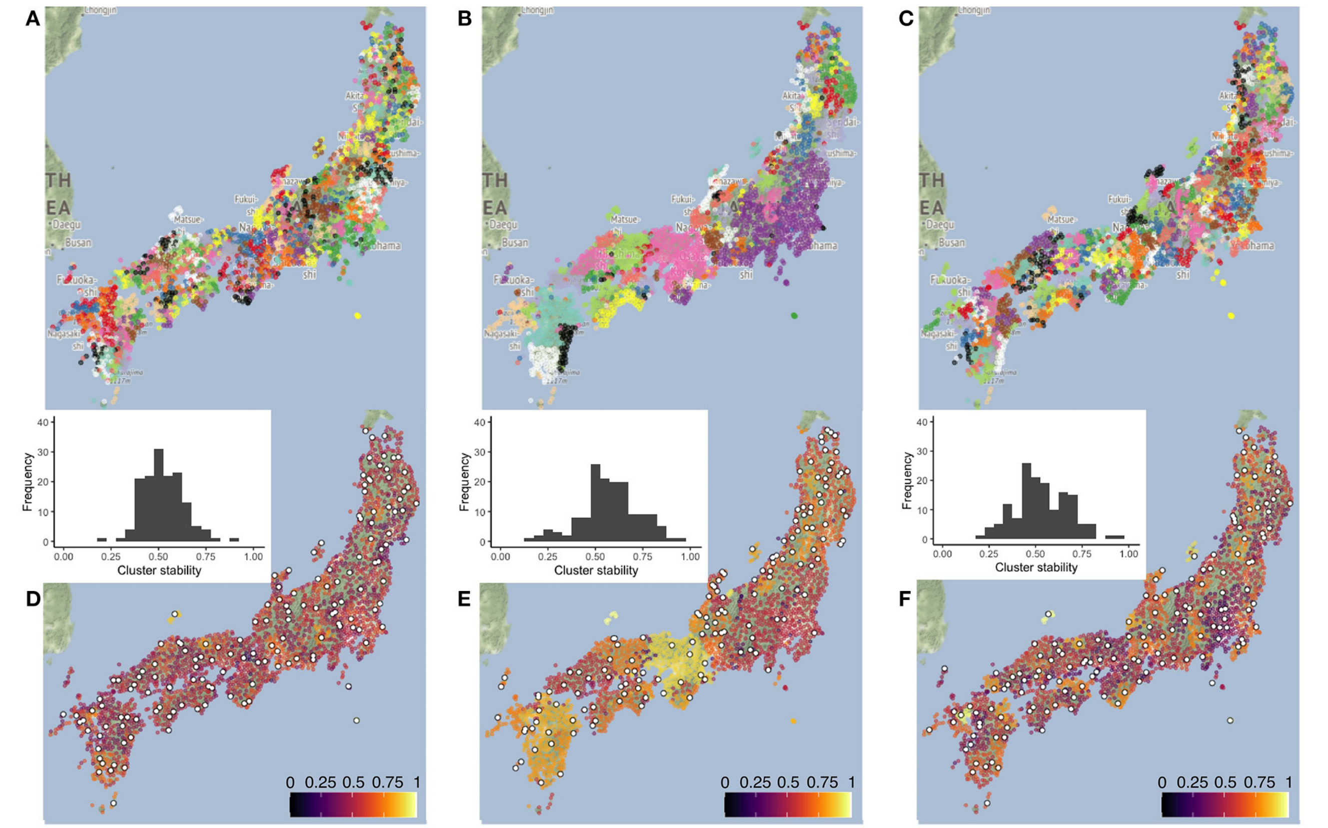


FIGURE 1 | The cluster maps **(A–C)**, stability maps **(D–F)**, and the stability histograms for the three clustering methods PAM (left), UPGMA (center), and Ward's (right) method, calculated for $k = 150$ clusters. Clusters in maps **(A–C)** are presented on a diverging scale of 15 colors, which repeat. Stability maps **(D–F)** also contain candidate survey sites selected based on linguistic centrality within their own cluster. Large differences are visible across the cluster solutions and smaller differences across the stability of the clusters. Besides, stability shows little spatial autocorrelation beyond the large size of clusters found by UPGMA.

calculated cluster and count the number of times each survey site becomes the central one. This approach is implemented in section 4.3.

Depending on the time elapsed since the original data collection, it might also be useful to estimate dialect change. One approach is to assume that "linguistic gravity" (Trudgill, 1974) has driven local varieties to become more similar to, e.g., the most populous nearby survey site. In this sense, linguistic gravity can be used to estimate language change emanating from local hubs into their *hinterlands*, making them more similar to the hub. Such patterns are often associated with dialect leveling, e.g., in Swiss German dialects (cf. Christen, 1998). This approach is also implemented on SDS in section 4.2.2 and section 4.3.2.

Geography playing a small role in selecting candidate sites is relatively small, as clustering happens in the linguistic dimensions. It is, nevertheless, intuitive to designate the spatially central point in a cluster as a candidate, and surveys in dialectology often set out from equidistant samples, based on thorough qualitative arguments. For example, in case of limited or biased available data, this strategy may be reasonable for the estimation of a hypothetical future linguistically central point.

TABLE 1 | Meilă's *VI* values, comparing the cluster partitions across the three bootstrapped clustering solutions for k = 20, 50, 100, 150, 200, 300, 400, and 500.

	k	UPGMA	Ward		*k*	UPGMA	Ward
PAM	20	1.9338	2.0656	**PAM**	200	2.4114	2.3403
UPGMA	20		1.4823	**UPGMA**	200		1.8859
PAM	50	2.1974	2.2349	**PAM**	300	2.1907	2.0897
UPGMA	50		1.6984	**UPGMA**	300		1.7146
PAM	100	2.4787	2.3981	**PAM**	400	2.0009	1.8902
UPGMA	100		1.9108	**UPGMA**	400		1.5331
PAM	150	2.5056	2.4467	**PAM**	500	1.8243	1.7581
UPGMA	150		1.9315	**UPGMA**	500		1.3471

The higher Meilă's VI, the more different the cluster partitions are.

Beyond these aspects and the objectives of the intended study, candidate survey sites might also be selected using external characteristics of the survey sites or a ranked eligibility measure of multiple characteristics. In case of studies interested in smaller areas or a few survey sites, qualitative methods may suffice from this point onward. If stable clusters are obtained, it is possible to investigate them one by one to choose the sites most appropriate for the contemporary dialectal variation and the intended study.

3.4.1. Candidate Selection in LAJ

For LAJ, we find linguistically central survey sites in each cluster by summing linguistic distances within clusters. The survey site with the smallest total linguistic distance within the cluster becomes the candidate site. In case of PAM, this point is exactly the *medoid*. **Figure 1** plots, for $k = 150$, the candidate sites for each clustering method in **Figures 1D–F**, as white points with a black contour. Candidate sites from PAM and Ward's method are identical in 63 cases, whereas their overlap with UPGMA is much lower (21 and 9, respectively).

Candidate sites in the case of UPGMA are distributed more evenly than the cluster structure, comprising several large clusters, suggests. This is due to singletons and unlikely clusters that are made up of several sites farther apart (such as the blue sites in Map B scattered within the largest purple cluster in the east—around Tokyo). Accepting these candidate sites as a reduced set of survey sites would cause problems in representation of spatially surrounding dialects.

The continuous nature of the data and the validity of FDP are confirmed by Jeszenszky et al. (2019). Based on the cluster structure, cluster stability patterns, and the patterns of candidate sites seen in **Figure 1**, we conclude that Ward's method is the most well-grounded for $k = 150$. Due to their spatially compact clusters it yields, Ward's method presents itself as the safest bet, knowing the bias in the data because of the phenomena it contains. In contrast, UPGMA produces unrealistically large dialect areas and unreasonable singletons, and PAM is less stable with more clusters overlapping in space.

3.5. Evaluation and Revision of Candidate Survey Sites

The candidate survey site sets resulting from the site reduction procedures are assumed to be representative of the original data. However, their main aim, as candidates, is to provide a quantitatively supported starting point for determining the sites that actually need to be researched. Several reasons call for a further qualitative evaluation of candidate survey sites. First, the linguistic basis of the site reduction might not be perfect due to various potential factors within the original database, the requirements of the intended study, and the circumstances that might have changed since the original survey. Second, dialect change may have progressed, due to people's changing way of life, mobility patterns, language attitudes, etc. Third, potential survey sites might have changed with regard to their sociodemographic settings, language policies, etc. Therefore, any set of candidate survey sites has to be revised in accordance with the requirements of the intended research, which potentially collects contemporary dialect data. Generally, the potential uncertainty about representativeness of contemporary dialectal variation increases with time elapsed since the original data was recorded, thus increasing the value of expert revision. Depending on the study's aims, the step of evaluation may result in swapping sites, adding sites that were originally not recorded, selecting more than one site from a cluster, rebalancing a clustering solution based on a spatial grid, etc. In section 4.4, we provide a qualitative revision of a candidate site set from SDS.

4. APPLICATION EXAMPLE: SDATS

In this real-life example, we present the entire site reduction procedure as applied to digital data from the "Sprachatlas der deutschen Schweiz" (SDS), with the aim of finding survey sites corresponding to the requirements of the contemporary dialect research project SDATS. Thus, the final goal is to find a way to represent the estimated contemporary variation, inferred from the original data and revised based on experts' field knowledge.

"Swiss German Dialects Across Time and Space" aims for candidate sites that are linguistically as different from one another as possible, thereby covering the largest swath of dialectal forms used. We carry out the clustering experiment with two different approaches on the same data set. First, Approach I is used for the demonstration of a generalizable methodology, presented in section 4.2. This approach applies the quantitative steps of the methodology similarly to the example in section 3. Second, Approach II is used to arrive at the survey sites actually used in SDATS, as detailed in section 4.3. This approach applies only the PAM clustering algorithm with a different custom-made validation approach. Then, candidate survey sites are revised to represent the contemporary dialectal variation, in section 4.4.

4.1. Linguistic Distance

Scherrer's digitized SDS database (termed Scherrer's data)[9] covers 289 linguistic variables: 107 phonetic, 118 morphosyntactic, and 64 lexical variables (Scherrer, 2021). SDATS's initial plans included revisiting 200 linguistic phenomena in SDS. At the time of selecting the survey sites, however, the extent of the overlap of SDATS variables with Scherrer's data was not clear yet, therefore all digitized variables were utilized for the site reduction.

The linguistic distance matrix is calculated similarly to section 3.2.1, based on Goebl's *Relative Identity Value* (RIV_{jk}) (Goebl, 1983; Jeszenszky et al., 2019). For each variable, the difference based on the variant categories is noted for each survey site pair, allowing for multiple answers. The final linguistic distance between a survey site pair is the proportion of the differing variables among those variables where an answer is present for both survey sites (n), or

$$D_{ij}^{ling} = \frac{\sum D_Q}{n} \tag{1}$$

where D_Q is the number of diverging variables regarding survey sites i and j. For example, if in survey sites i and j answers for all linguistic variables are in different variant categories, then a linguistic distance of 1 is assigned to this survey site pair[10]. To

[9] The digitized data, together with its documentation is available in a tabular format at dialektkarten.ch, where individual variables are also interactively mapped.

[10] For more details, see Jeszenszky et al. (2019, p. 8–9).

discover linguistic distances in an aggregate manner, it is possible to use multidimensional scaling and thematic mapping[11].

4.2. Site Reduction: Approach I—Bootstrap Clustering

"Swiss German Dialects Across Time and Space" aims to select 125 survey sites and has the objective of collecting a balanced set of phenomena across linguistic levels Therefore, we intend to use data from SDS such that is also balanced across the linguistic levels. We group the linguistic variables according to the linguistic levels and calculate the linguistic distance matrices for each of them. To counter the higher numbers of morphosyntactic and phonetic variables, the mean value of these three matrices (termed the *mean linguistic distance matrix*, $-\overline{LD}$) is the input for the clustering steps. Note, however, that doing so leads to the increased weight of individual lexical phenomena.

We apply the three clustering methods presented in section 3.3.1. We perform clustering with bootstrapping on $\overline{LD}$ using PAM, UPGMA, and Ward's method from the `fpc` package, with $k = 125$, in accordance with the hard criterion in SDATS. Similar to section 3.3.1, we use the stability values associated with clusters as the method of internal validation. We also calculate Meilă's *VI* to compare clustering solutions' similarity across methods and across clustering solutions resulting from different subsets of the data. In addition, in section 4.2.2, we test how well different sets of candidate survey sites represent the original $\overline{LD}$.

4.2.1. Clustering and Validation

Figure 2 maps cluster solutions based on the three clustering methods, PAM (A), UPGMA (B), and Ward's method (C). As expected, cluster members are also spatially clustered in the overwhelming majority of the cases. In a few cases, members of a cluster are separated by members of other clusters. In addition, singletons are present. Both PAM's and Ward's maps show spatially compact clusters (corresponding to the FDP), while the UPGMA map is more prone to producing larger clusters and clusters of singleton outliers. UPGMA finds 50 singletons, while PAM and Ward's method find 27 and 17, respectively. These patterns are structurally similar to the clustering results of LAJ.

Figures 2D–F present the stability of clusters. It is visible, here, that some clusters are stable regardless of the clustering method, while values vary in other areas. Most of the Swiss Plateau[12] shows low cluster stability, especially with PAM. Overall stable regions include the cantons of Schwyz (SZ), Uri (UR), Obwalden (OW), Glarus (GL), the Entlebuch region in the canton of Lucerne (LU), the Haslital region and the SE part of the Bernese Oberland in canton Berne (BE), and the Eschenbach region of canton St. Gallen (SG). The singleton survey sites, e.g., in the canton of Graubünden (GR) and elsewhere do not show very high stability, independent of clustering method. Interestingly, UPGMA and Ward's method provide stable clusters in the canton of Basle-Country (BL), while PAM and Ward's method show stability in the Oberland region of canton Zurich (ZH).

Stability values are also presented as histograms in **Figures 2G–I**. The skew toward the right implies that the bulk of clusters are stable, with little difference between the clustering methods. Based on the stability values, the cluster structures and the field expertise of SDATS project members, each cluster solution is deemed acceptable for the production of *candidate* sites. However, there are some evident drawbacks. PAM's stability, on average, seems lower, but the differences in the maps and histograms are visually not as substantial as those seen in the application to LAJ data. UPGMA's larger clusters and singletons are often linguistically not supported (e.g., an expert would expect to find more clusters in the canton of Valais—VS). Finally, the compact and similar-sized clusters of Ward's method are tempting for dialectology, but they are often unreasonable, e.g., in the Swiss Plateau.

4.2.2. Selection of Candidate Survey Sites

The next step in the methodology is appointing a candidate survey site within each cluster. We can select linguistically central sites, defined by the smallest total linguistic distance toward cluster members. Appointing this site intrinsically makes PAM a practical method for the application. However, as SDATS aims to investigate contemporary dialectal variation, we select candidate survey sited based on estimated potential dialectal change since data collection in SDS. We aim to find sites that have potentially influenced their local surroundings since 1950, assuming, based on Trudgill's linguistic gravity theory (1974), their surroundings have become more similar to them (Christen, 1998; Szmrecsanyi, 2012; Schmid et al., 2019). To address this, we select survey sites with the highest population in 2018 from each cluster, using official census data (BFS, 2018).

Figure 3 presents these two kinds of candidate survey sites sets for the three clustering methods (A—PAM; B—UPGMA; and C—Ward's). Linguistically central sites are depicted by +'s, and sites with the highest population by ×'s. In Map D, all candidate sites from the other three maps are stacked, to show the potential eligibility of any SDS survey site. In the case of UPGMA (Map B), the two requirements overlap in more than half of the cases, though this happens less frequently for the other two clustering methods. Maps A–C convey the message that the site with the highest population might not be the linguistically central or representative site with regard to the original data, suggesting that estimating future linguistic scenarios based on linguistic gravity should be approached with caution. In Map D, overlaps of the symbols show a higher potential eligibility of sites in the Alps, especially in the canton of Graubünden (GR), with the latter due to the high proportion of singleton clusters. This, nevertheless, hints at the presence of unique dialects.

To evaluate these candidate site sets, we test if they are representative of the original survey site set of SDS (573 sites). Technically, we test if the similarity of the distributions in the linguistic distance matrices of the candidate set and the SDS set ($\overline{LD}$) is statistically significant. Since the values in the matrices of the original set and in the candidate sets are not

[11] Kellerhals (2014) has already produced the MDS maps for each linguistic level, and as an aggregate, based on the contemporary status of Scherrer's data. Also consult Yves Scherrer's homepage, dialektkarten.ch, for average linguistic maps and other parameter maps based on several linguistic atlases.

[12] *Germ.:* "Mittelland," the relatively flat part of German-speaking Switzerland from Lake Constance to Lake Bienne in the west.

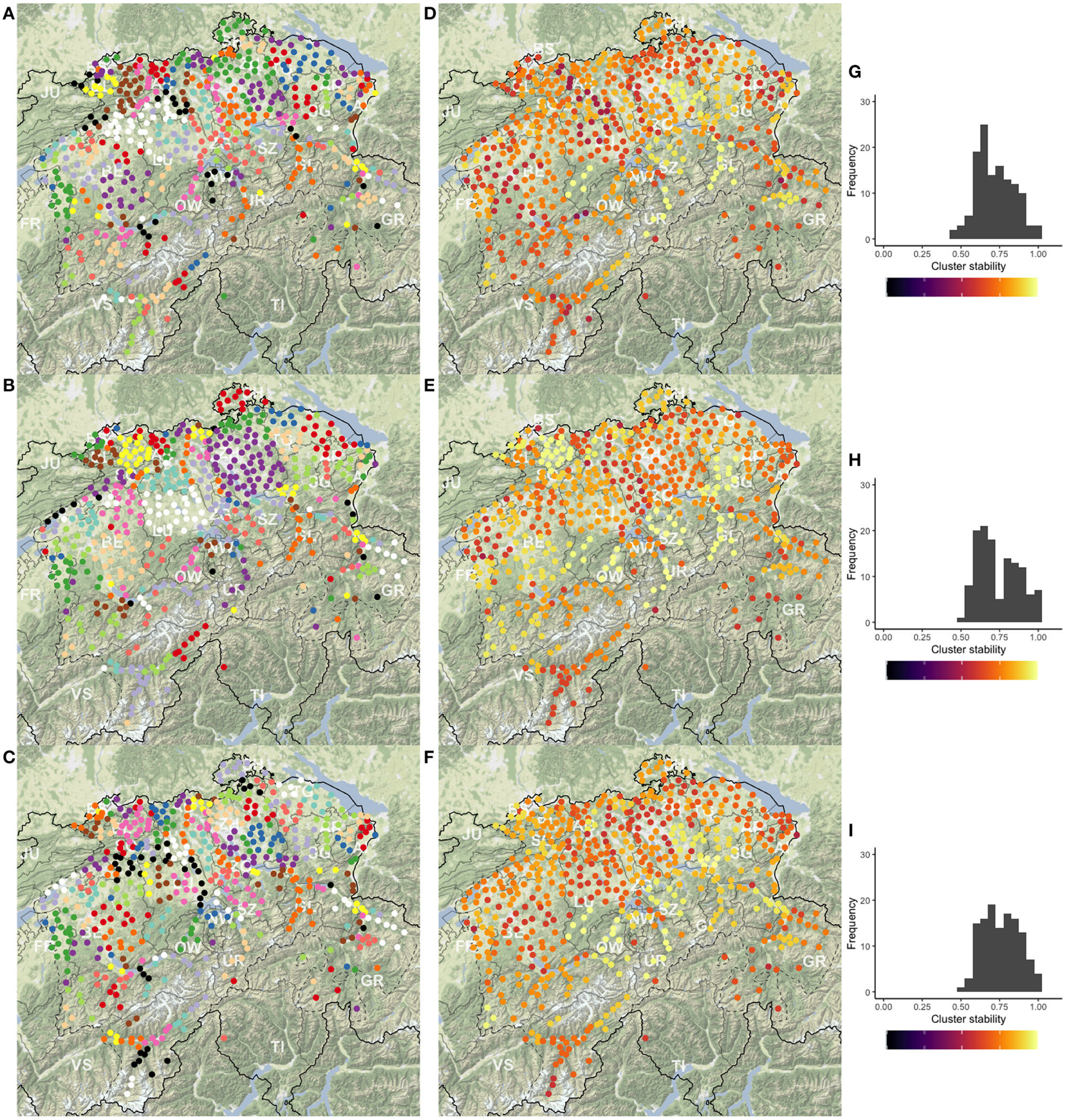


FIGURE 2 | The cluster maps **(A–C)**, stability maps **(D–F)**, and the stability histograms **(G–I)** for the three clustering methods PAM (top), UPGMA (center), and Ward's (bottom) method, calculated for k = 125 clusters. Clusters in maps **(A–C)** are presented on a diverging scale of 15 colors, which repeat. Substantial differences are visible across the cluster solutions and smaller differences across the stability of the clusters across the methods. Importantly, clusters in some regions stay stable independent of the clustering method.

normally distributed, we use the *Kruskal—Wallis test* to test the significance of the differences. Affirming this, the *pairwise Wilcoxon rank sum test* allows us to test which candidate sets' linguistic distance matrices have a significantly different distribution from the original $\overline{LD}$. In addition to the candidate sets, we test the performance of random site sets as well. For each of the clustering methods, we create ten random site sets, selecting one random site from each of their clusters. Further, we create 1,000 unrestricted random site sets from the SDS survey sites.

The Wilcoxon rank sum tests shows that no candidate site set presents a significant difference from the original linguistic

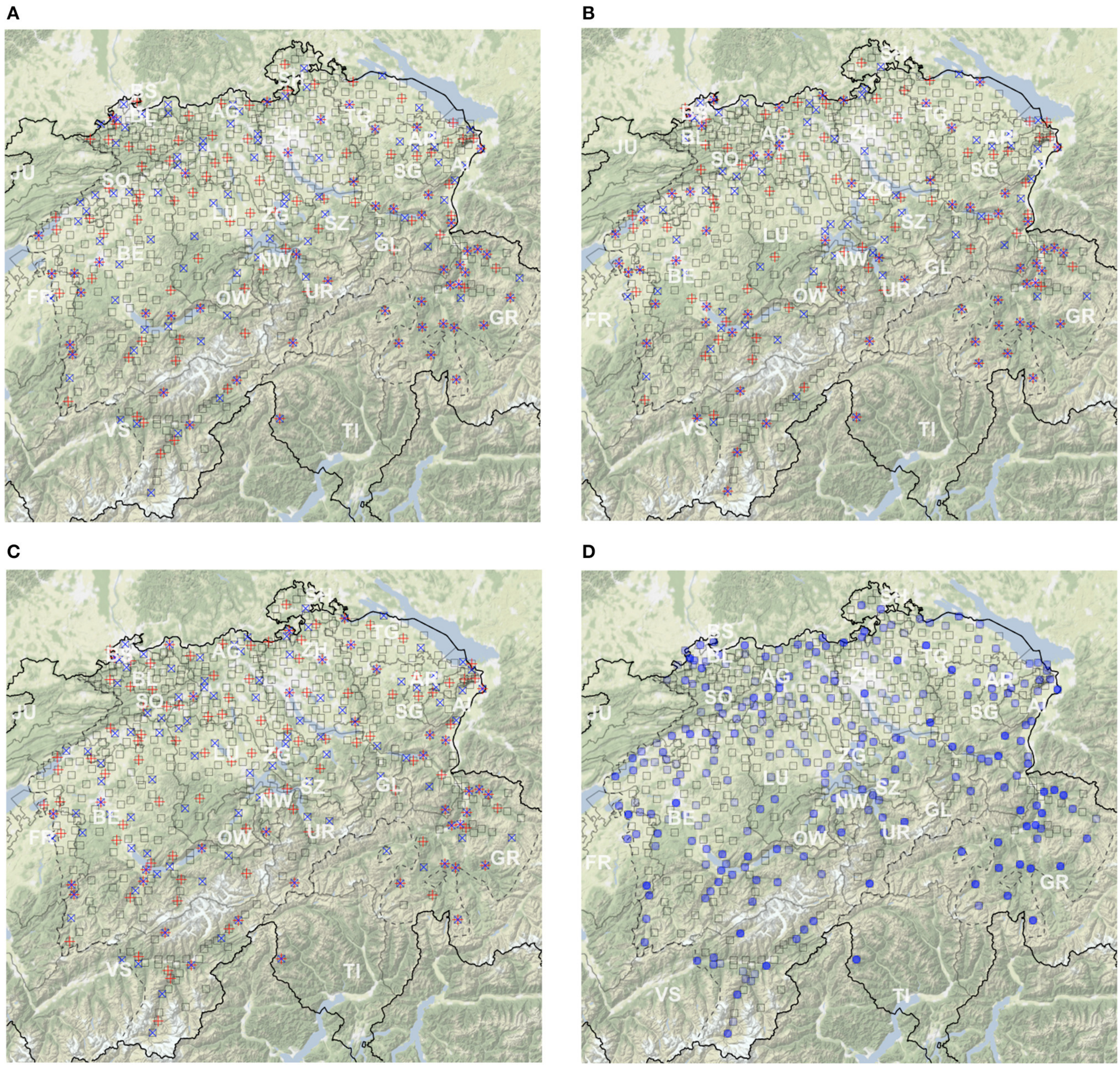


FIGURE 3 | Maps of the two types of candidate survey sites. Centers are appointed by the smallest total linguistic distance within a cluster (red +), and the highest population within a cluster (blue ×), for each clustering method (**A**—PAM; **B**—UPGMA; and **C**—Ward's method). Map **(D)** shows all candidate survey sites in a stacked manner. All maps contain the original SDS survey sites in the background (gray squares).

matrix set. Importantly, however, only 27.64% of the unrestricted random sets show a significant difference from $\overline{LD}$. This value is still 40.7% when sinking p-value's threshold to 0.001. At the same time, random site sets from clusters never exhibit a p-value over 5×10^{-13}. Thus, there is substantial possibility that an unrestricted random sample becomes representative of the whole population. We argue that this is due to sampling one out of five points, a relatively large sample, and that a threshold of representativeness has to be cautiously applied by the researcher.

Figure 4 presents some distributions of the linguistic distance matrices of the candidate site sets in relation to the original $\overline{LD}$. **Figure 4A** presents six random sets from the previous test, with random sets from clusters in brown and unrestricted random sets in yellow. Brown lines stay below the distribution of $\overline{LD}$ in the left side and overshoot $\overline{LD}$ at its peak. Yellow lines follow the distribution of $\overline{LD}$ more exactly, but this is not always enough to be representative. **Figure 4B** presents the densities of the candidate site sets. All lines stay somewhat below the distribution of $\overline{LD}$ on the left side and overshoot the $\overline{LD}$ at its peak. This means

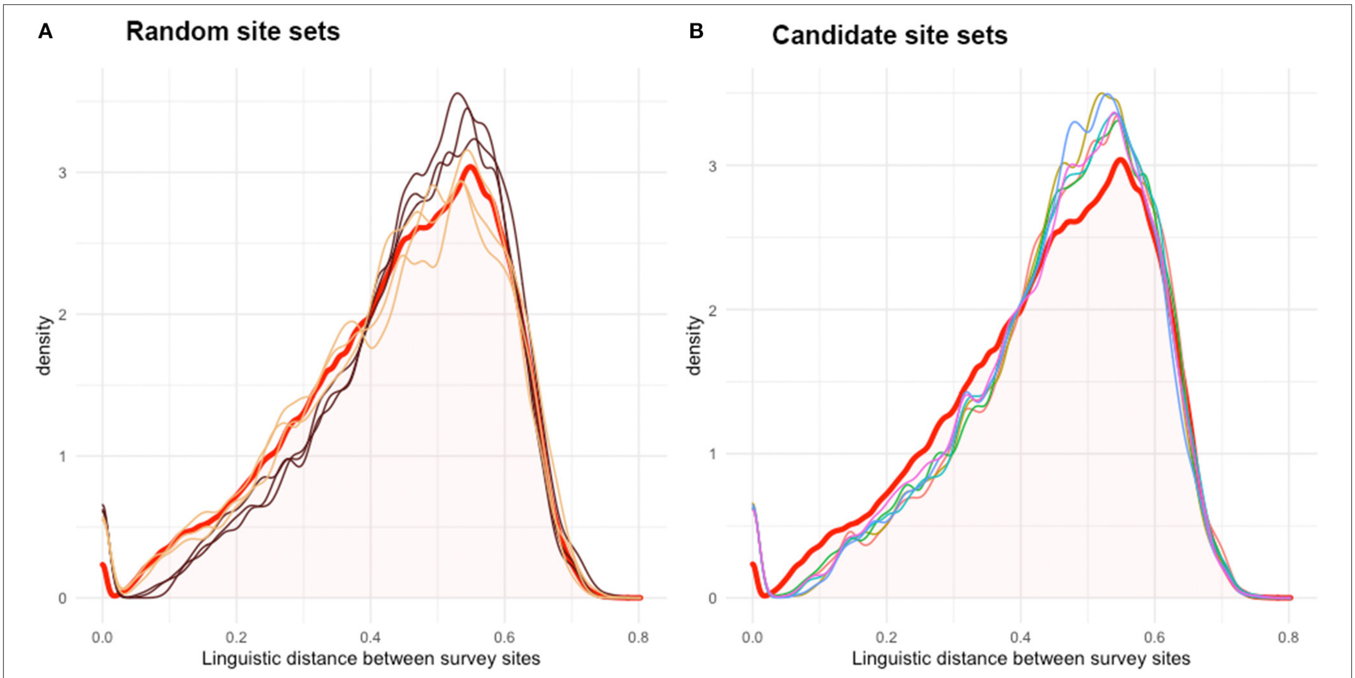


FIGURE 4 | Density graph of the values within linguistic distance matrices. In **(A)**, the linguistic distance matrices from three unrestricted random sets (yellow) and random sets from the resulting clusters of each clustering method (brown) are plotted together with the original mean linguistic distance matrix ($\overline{LD}$—red). In **(B)**, linguistic distance matrices from the candidate survey site sets are presented. Beside unrestricted random sets, all sets overshoot the peak density of $\overline{LD}$. Thus, on average, linguistic distances are larger within these candidate site sets, as expected, and their members are linguistically more different from one another.

that the candidate survey site sets include more of those survey sites that have a higher linguistic distance toward one another, ultimately intensifying the variation present in the candidate set while overlooking survey sites that are less diverse, thus less different from one another.

4.3. Site Reduction: Approach II—Candidates Resulting From Ranking

This section serves the purpose of detailing the site reduction approach implemented to define the conclusive candidate site set for SDATS, which is used from section 4.4 for the qualitative revision. We present a customized methodology of cluster analysis and site selection to fulfill two aims. First, we address SDATS' requirement of balance across linguistic levels. Second, we address the requirement of inferring a future linguistic situation based on the theory of linguistic gravity. To this end, we use a special-purpose cluster validation technique and build the qualitative requirements of the SDATS project partly into the clustering step (i.e., selecting a candidate site with a relatively high population from a cluster). We arrive at the candidate survey site set by ranking the survey sites based on two measures introduced below, one related to their stability in their clusters (J), and another based on their population (P_{top}).

Scherrer's data are imbalanced across linguistic levels and it contain 107 phonetic, 118 morphosyntactic, and 64 lexical items. In section 4.2, we calculated the mean linguistic distance based on the three linguistic levels. This means, however, that the weight of each lexical item is almost double the morphosyntactic items. The approach introduced here is proposed as an experimental method to counter this effect. In order to get a sample of items representative of each linguistic level, we create S subsets, drawing equal numbers of random items from each of the three linguistic levels. On the one hand, we randomly select 64 items from each linguistic level (referred to as subsets S^{64})[13]. On the other hand, we randomly select 20 items from each linguistic level (referred to as subsets S^{20}). The number 64 is decided by the number of lexical items in Scherrer's data, the lowest among the linguistic levels. In parallel, sets of 20 items are used to decrease the bias assumed to be caused by the constant presence of all 64 lexical items in the S^{64} subsets. We create 35 subsets of S^{20} (S^{20}_{101}, S^{20}_{102}, S^{20}_{103} ... S^{20}_{135}), and 30 subsets of S^{64} (S^{64}_{1}, S^{64}_{2} ... S^{64}_{5}, and S^{64}_{201}, S^{64}_{202}, S^{64}_{203} ... S^{64}_{225})[14].

The overlap of items across random subsets is visualized in **Figure 5**. It is visible that the overlap is much smaller (warmer, reddish colors) among S^{20} subsets, compared to S^{64} subsets, around 20% on average, with some outliers. Overlaps across S^{20} and S^{64} subsets are much larger (colder colors), with an average of around 70% overlap. The overlaps among S^{64} subsets are more uniform (around 70%, with less deviation), as they include by

[13] We use specific *seeds* in R to create reproducible randomized subsets. Setting a seed determines the starting number used to generate a sequence of random numbers; using the same seeds ensures the reproducibility of the same subsets.

[14] Seeds of S^{20} and S^{64} subsets (shown as the lower indices) do not overlap to avoid complete overlap across the items selected.

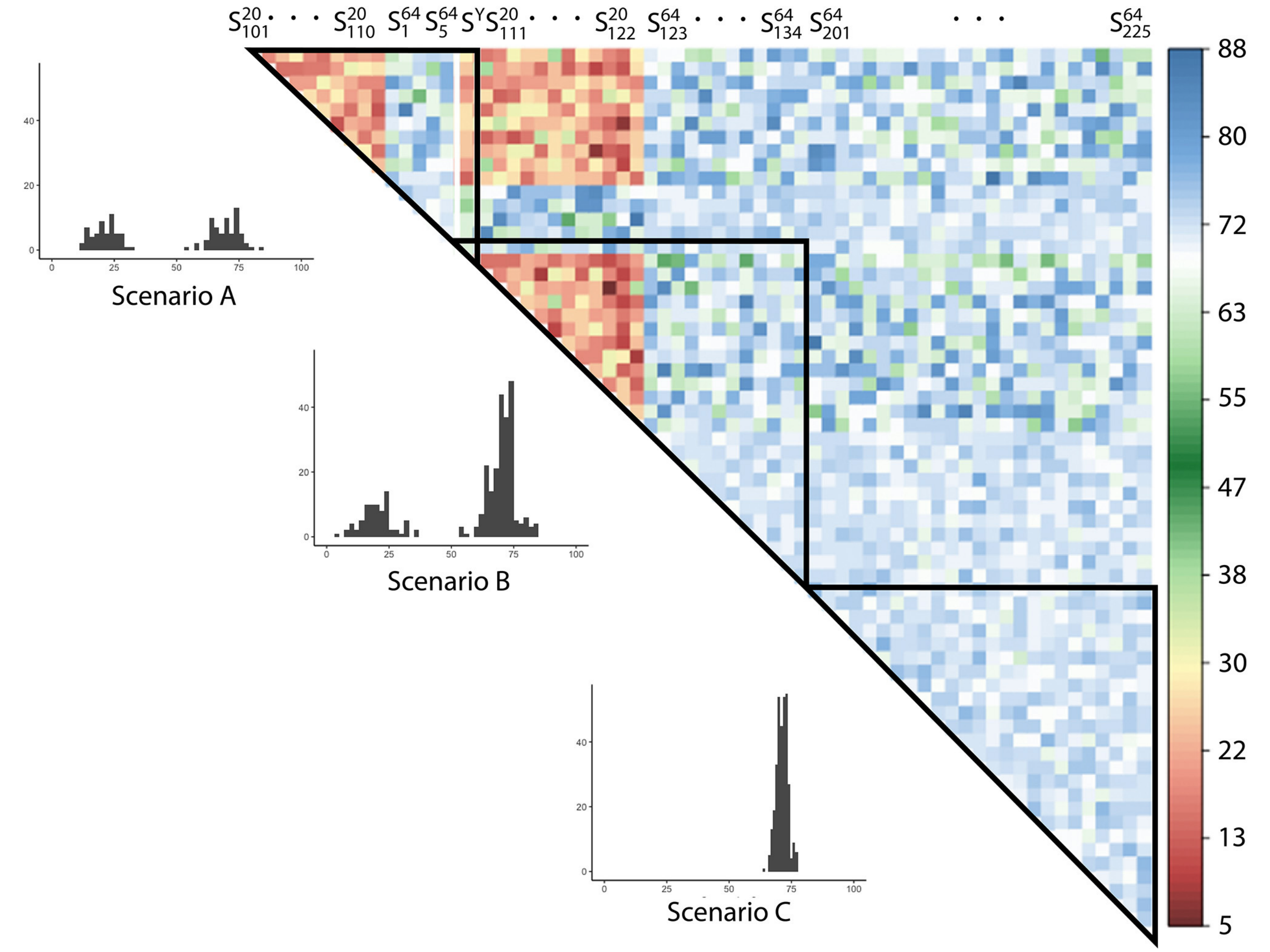


FIGURE 5 | The number of SDS linguistic items overlapping across balanced random subsets. Group-internal overlaps belonging to Scenarios A–C (introduced in section 4.3.2, below) are highlighted in triangles. The histograms show the amounts of overlap for the three scenarios.

definition all 64 lexical items and the majority of phonetic and morphosyntactic items.

For each *S* subset, we calculate the linguistic matrices. Effects of random item selection in the subsets are shown by Pearson correlation coefficient values across their linguistic matrices, which are almost always above 0.9, with the lowest values around 0.75. The high values ($R^2 \geq 0.62$) confirm the similarity across the random subsets even in cases of smaller item overlaps across S^{20} subsets.

4.3.1. Clustering and Validation

We carry out PAM clustering with $k = 125$ on the linguistic distance matrix calculated from each of the *S* subsets, using the `cluster` package (Maechler et al., 2019) in R. **Figure 6** shows the clusters resulting from PAM runs based on variables in five subsets. Structurally, the cluster patterns look similar to the PAM map (Map A) in **Figure 2**.

To justify using PAM, we test the similarity of *S* subsets' cluster solutions to the cluster solutions of $\overline{LD}$ (PAM, UPGMA, and Ward's method), using Meilă's *VI*. The most important results are shown in **Table 2**. It becomes visible that PAM clusterings of *S* subsets are more similar to the PAM clustering of the mean linguistic distance matrix ($\overline{LD}$_*PAM*) than the Ward's ($\overline{LD}$_*Ward*) or UPGMA clustering of $\overline{LD}$ ($\overline{LD}$_*UPGMA*). Therefore, if we accept that each of the clustering methods produce linguistically plausible cluster partitions on $\overline{LD}$, then the PAM clustering of the *S* subsets can also be accepted with a high probability.

We validate clusters using a custom method resembling the noisy clustering and bootstrapping approaches often used for cluster validation in dialectometry. For each survey site pair, we note the number of occurrences when the two sites are clustered together. Then, for each survey site, we calculate the percentage (termed *J*) of clustering runs, in which the site is clustered together with the same other survey sites. If survey sites *h*, *i*, and *j* always fall into the same cluster and there is no other survey site ever falling into this cluster, then each of the sites *h*, *i*, and *j* get the maximal *J* value. A survey site that always becomes a

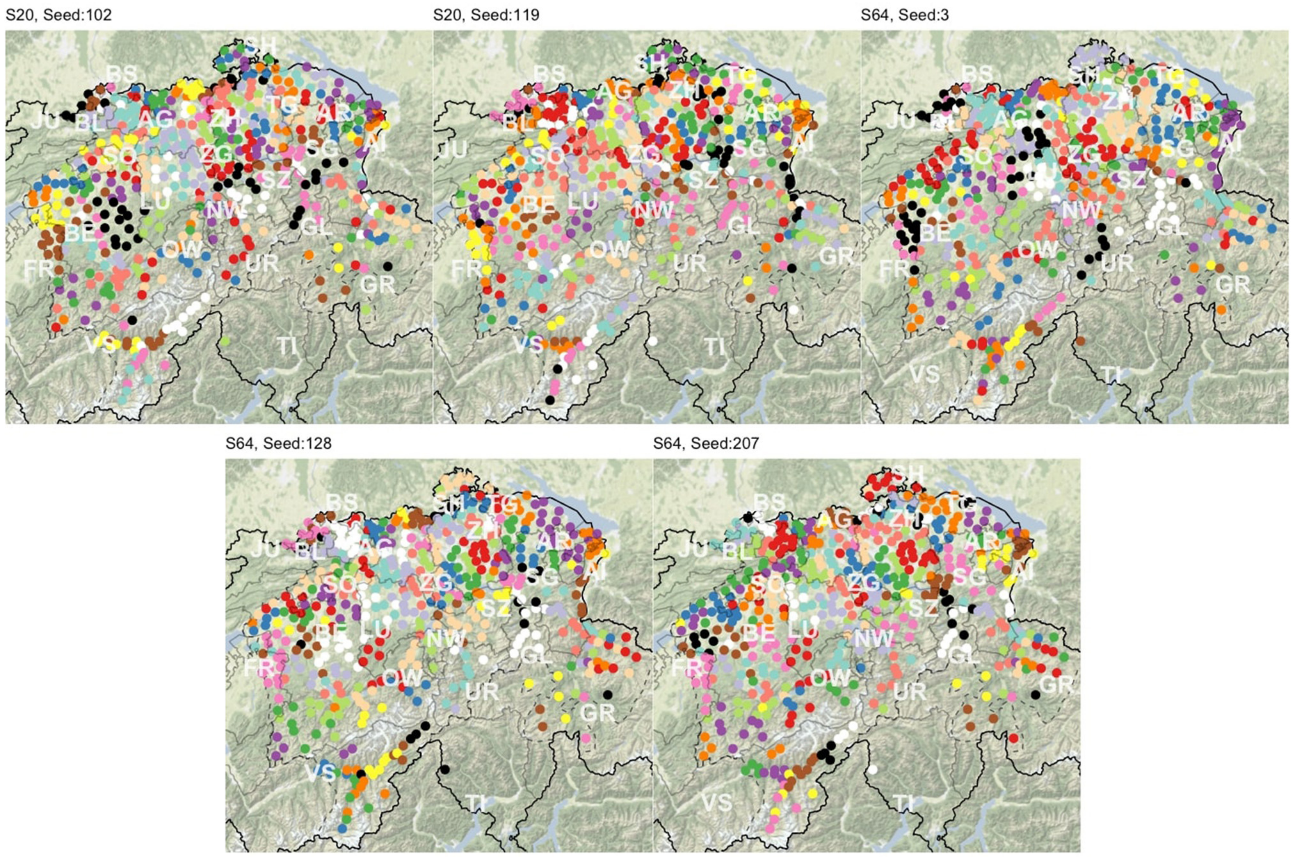


FIGURE 6 | Clusters resulting from five balanced random subsets of SDS variables. Clusters are presented on a diverging scale of 15 colors, which repeat.

singleton would also get this value, giving the chance for very local but unique dialects to stand out as stable clusters.

4.3.2. Selection of Candidate Survey Sites

As in section 4.2.2, we aim to find those survey sites that most affected their surroundings in the last 70 years through the effect of linguistic gravity. To this end, we find the survey site with the highest contemporary population (BFS, 2018) in each cluster, for each of S subsets' cluster solution. For each survey site, we note the proportion of S subsets' cluster solutions at which the survey site exhibits the highest population in their own cluster. This proportion is termed P_{top}.

Survey site eligibility is then ranked based on J and P_{top}. The scatterplots in **Figure 7** show these factors that test the correspondence of SDS sites to SDATS requirements. The x-axis, along with the point color and size, presents P_{top} (green, over 50%; blue, between 25 and 50%; and gray, below 25%, the latter corresponding to a low eligibility for the final SDATS set). The y-axis, along with the background color, shows the J value of the site, running from dark purple (low "stability in own cluster") to yellow (high "stability in own cluster"). An ideal survey site would score high with regard to both requirements, reaching the top right corner of the scatterplots. Based on the point color, size, and background color, J and P_{top} can be transferred to the maps on the left. Values of J are shown in Maps A–C as the colors of Voronoi-polygons around their SDS sites, illustrating the areal distribution of J. Because clusters mostly contain more than one survey site, similar J values are expected to cluster in space.

Before we select the candidate sites based on the clustering solutions of all subsets, however, we revisit the potential bias caused by the imbalance across linguistic levels. In the data set, lexical variables make up the smallest portion. However, lexical variables are the most diverse, therefore their variation patterns are potentially the most different from one another and, thus, are associated with greater linguistic distances. **Table 3** presents the mean, median, SD, and variance values of the three linguistic levels, with values of the lexical level substantially exceeding the other two levels.

As we select roughly one-third of the lexical variables in S^{20} subsets, there will be a variation in the effect across subsets (as deductible from the overlaps across S subsets in **Figure 5**). S^{64} subsets, however, contain all lexical variables, always conveying the full effect of the lexicon.

We aim to select the SDATS survey sites based on a balance across linguistic levels. In order to assess the effect of lexicon, we set up three Scenarios which pool the cluster solutions from a number of S subsets. The difference between Scenarios is the

TABLE 2 | Meilă's *VI* values, comparing the cluster partitions across linguistic distance matrices of *S* subsets and cluster partitions of $\overline{LD}$ using the three clustering methods, PAM, UPGMA, and Ward's method.

	Clustering A	Clustering B	Meilă's VI
1	S^{64}_{224} PAM	$\overline{LD}$_PAM	0.3796
2	S^{64}_{126} PAM	$\overline{LD}$_PAM	0.4069
3	S^{64}_{218} PAM	$\overline{LD}$_PAM	0.4096
4	S^{64}_{217} PAM	$\overline{LD}$_PAM	0.4288
5	S^{64}_{4} PAM	$\overline{LD}$_PAM	0.4303
6	S^{64}_{206} PAM	$\overline{LD}$_PAM	0.4370
7	S^{64}_{129} PAM	$\overline{LD}$_PAM	0.4479
8	S^{64}_{211} PAM	$\overline{LD}$_PAM	0.4485
9	S^{64}_{208} PAM	$\overline{LD}$_PAM	0.4509
10	S^{64}_{204} PAM	$\overline{LD}$_PAM	0.4583
...			
46	S^{64}_{128} Ward	$\overline{LD}$_PAM	0.6302
47	S^{64}_{215} Ward	$\overline{LD}$_PAM	0.6398
48	S^{64}_{4} Ward	$\overline{LD}$_PAM	0.6473
49	$\overline{LD}$_Ward	$\overline{LD}$_PAM	0.6476
...			
76	$\overline{LD}$_UPGMA	$\overline{LD}$_PAM	0.7366
77	S^{64}_{221} Ward	$\overline{LD}$_PAM	0.7424
78	S^{64}_{130} UPGMA	$\overline{LD}$_PAM	0.7434
...			
143	phon_Ward	$\overline{LD}$_PAM	0.9909
...			
160	phon_PAM	$\overline{LD}$_PAM	1.0541
...			
255	...	...	...

proportion to which they contain S^{64} subsets, those that convey the full effect of lexicon:

- In Scenario A, the proportion of the full effect of lexicon is 1/3,
- In Scenario B, the proportion of the full effect of lexicon is 1/2,
- Scenario C is entirely made up of S^{64} subsets, thus always conveying the full effect of lexicon.

Within the Scenarios, we employ a consensus approach based on the numerous cluster solutions pooled, expecting the cluster solutions, which are somewhat different across subsets, to converge toward a central value. S^{64} subsets overlap to a larger degree than S^{20} subsets, causing cluster solution across S^{64} subsets to be more similar. Therefore, the higher the involvement of S^{64} subsets, the higher J values are expected. The difference of J values across the maps and scatterplots in **Figure 7**, thus, demonstrates the effect of imbalance across linguistic levels.

With the qualitative revision already taking a foothold in the cluster validation steps, the initial candidate survey sites are selected based on their ranking of J and P_{top}. Scenario C's map, regarding the J values, resembles the stability map of PAM (Map A) in **Figure 2** (despite J values not being equivalent to the stability values in a bootstrapping approach), which hints at the similarity of Scenario C and the bootstrap clustering in section 4.2.1. Although our original aim was countering the overweight of lexicon present in $\overline{LD}$, the candidate survey sites resulting from the customized site reduction do not differ substantially across the three Scenarios. This is indicated by the set of survey sites highlighted with their names in Scenario C's map, which forms a superset of sites visible in Maps of Scenario A and B. Thus, it is visible that the qualitative decision of using P_{top} as a candidacy factor overwrites the effect of linguistic levels. Still, the potential effect of linguistic imbalance is a valid limitation for applications of the general methodology.

Initially, based on the rankings in J and P_{top}, 114 candidate survey sites are selected, fewer than the number of clusters originally sought. This is a result of a manual intervention, which is due to the fact that for the aims of SDATS, it is not the number of clusters or their identity that is relevant, but the survey sites' ranking on J and P_{top} values. Beyond the first 114 survey sites, further sites' ranking with regard to either J or P_{top} was too low, thus we decided to leave it to the qualitative revision to fill up the selection, as we ultimately maintain the aim of selecting 125 survey sites.

4.4. Revision Based on Linguistic and Sociodemographic Factors

For the qualitative revision of candidate survey sites, we use the candidate site set resulting from section 4.3. This means that the candidate set of survey sites is not equal to the number of clusters sought in the earlier steps. Nevertheless, qualitative revision is not necessarily bound by the clusters or candidate sites yielded by the quantitative steps.

Several reasons impede us from relying fully on the clustering results. First, the cluster partitions reflect the state of the dialectal landscape around 1950, in contrast to the SDATS requirements of investigating contemporary local colloquial dialects. Second, Switzerland has undergone sociodemographic changes, often affecting the composition of the population in settlements. People have become more mobile, and the communities in certain towns and villages recorded in SDS might have changed massively due to industrialization, urbanization, and suburbanization. Third, the digital linguistic data are not entirely optimal for the site reduction. Even if the 289 items in Scherrer's data were representative of SDS, the categorization of the variants within an item corresponds to the needs of SDATS for only 45 items. Besides, if each item is supposed to have the same weight in the process, neither Approach I nor Approach II can completely preclude the disproportionate effect of linguistic levels.

To address these factors, the initial candidate survey site set is evaluated from the following viewpoints, including the indispensable insight of linguists with expertise in past and contemporary dialectal variation. We inspect:

- whether important sociodemographic changes could have occurred at the candidate sites leading to a change or a mixture of dialects. If there was a remarkable change (such as a population boom due to extraordinary economic prosperity, or becoming a touristic hotspot), the location was eliminated from the list of candidate sites (e.g., Uster, canton of Zurich—ZH, or Klosters, canton of Graubünden—GR);

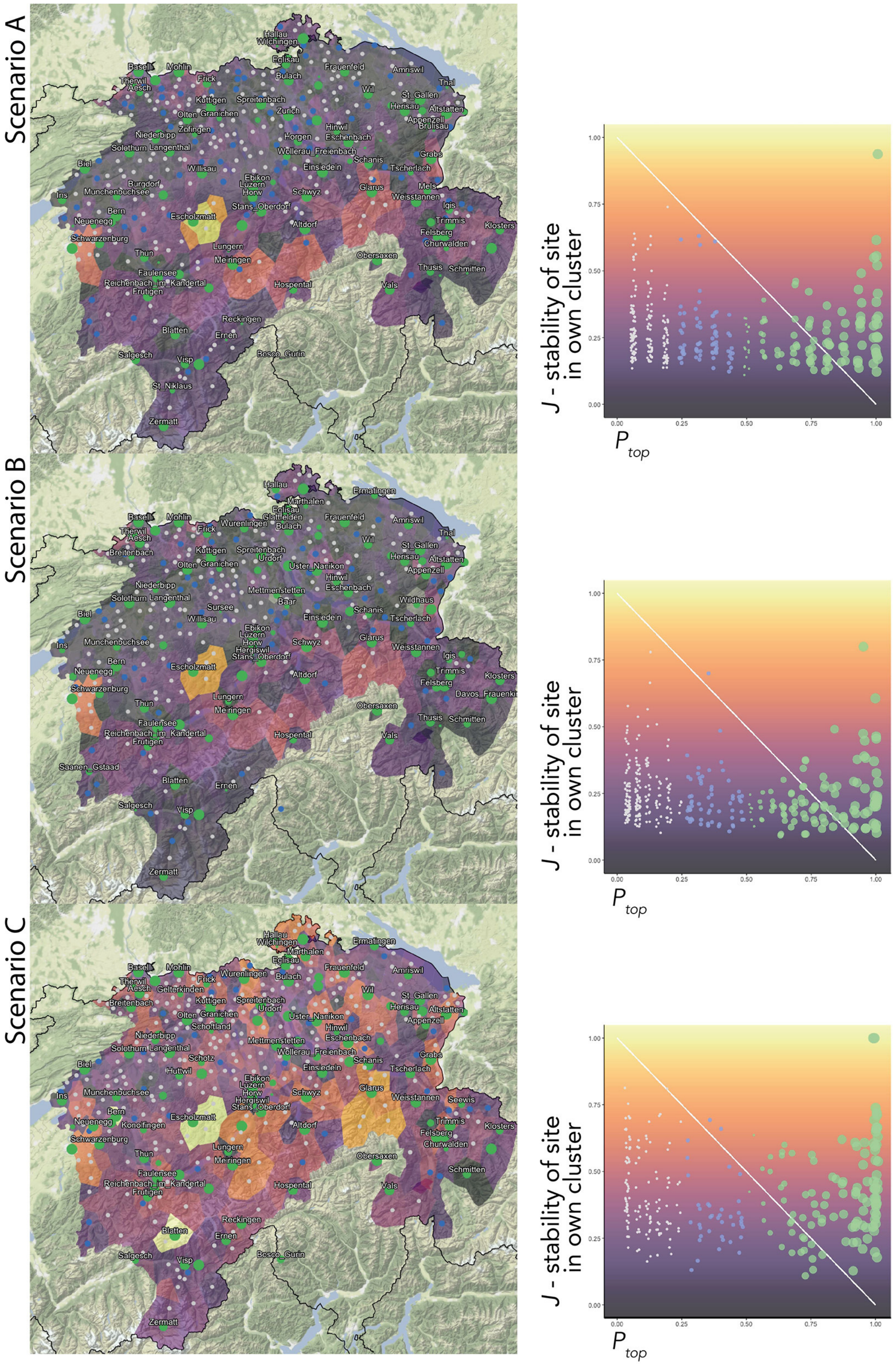


FIGURE 7 | Composite maps and graphs presenting the two candidacy factors deciding the ranking of an SDS survey site to become a candidate site for SDATS. The proportion of occasions a survey site was clustered together with the same others, *J* (stability of a site in its own cluster) is mapped between dark purple and yellow hues, where yellow means higher stability. The number of times (among the clustering solution on different subsets) a site has the most inhabitants in its own cluster is shown by P_{top}. The best candidate sites score high in both P_{top} and *J*. Such sites are presented as green circles on lighter polygons in the map.

TABLE 3 | Descriptive statistics of each linguistic level's linguistic distance matrices.

	Lexicon	Morphosyntax	Phonology
Mean	0.5337	0.4013	0.4001
Median	0.5652	0.4158	0.4216
Standard deviation	0.1771	0.1505	0.1354
Variance	0.0314	0.0226	0.0183

Besides higher linguistic distance values on average, lexicon has a wider statistical spread as well.

- whether a candidate site's location has merged into another community (e.g., Masans into Chur, GR);
- whether a community is very small or has lost many inhabitants, creating difficulties for recruiting enough respondents from all social backgrounds (e.g., we removed Weisstannen, canton of St. Gallen—SG, but kept Hospental, canton of Uri—UR);
- whether it is known that the local dialect is remarkably interesting (for the general public or from objective linguistics viewpoints). In some cases, local studies have documented change and peculiarities, validating the candidacy of some potential sites (e.g., Bosco Gurin, a partly German-speaking village in the Italian-speaking canton of Ticino—TI, or Blatten, a village in the secluded Lötschental valley in the canton of Valais—VS);
- whether the candidate site is perceived as linguistically representative of the region. For example, the city of Basel is traditionally not regarded as representative of its surrounding region;
- and, less importantly, whether there is a chance for an equidistant choice of survey sites. Following traditional sampling in dialectology, we might select a survey site that makes the survey site set equidistant (counterexamples include candidate sites separated by linguistically significant cantonal borders, e.g., between Niederbipp—BE, and Oensingen, canton of Solothurn—SO).

4.4.1. Revision by Dialectological and Sociogeographic Expertise

Following the aspects listed, the candidate survey sites are revised. The revision is done partly based on sites' rankings on J and P_{top}: when a candidate has to be removed, we often turn to these rankings for the next candidate or to validate the choice made based on other factors. For example, Lucerne, Horw, and Ebikon (LU), each were included in the initial set of 114 candidates. However, their geographic proximity allowed them to became a city complex in the last half century, functioning essentially as one unit, with a potential to drawing population from all over Switzerland. Therefore, we only chose Lucerne, the center, to represent this complex. Further, SDS survey sites that have merged in the last 70 years are treated as one, such as Schwamendingen and Zürich (ZH), or Masans and Chur (GR).

Next, we amended the candidate list in a qualitative fashion based on dialect expertise of the SDATS project members. For example, we added locations with local research already present and deemed interesting, such as Jaun (canton of Freiburg—FR), a German-speaking isolate, and Obersaxen (GR), a former island of Walser dialect. We also strived to cover some local dialects deemed peculiar due to the dwindling German-speaking population (Bosco Gurin, TI) or isolation (Vättis, SG). Importantly, major cities and towns of central importance have been taken into the SDATS sample regardless of the clustering results.

At this stage, external dialect experts made further suggestions about available SDS survey sites that are overall representative in their region today, thus not necessarily reflected in the digital data. Some external dialect experts objected to the involvement of urban centers due to the assumptions that urban mixed dialects have already been outliers in SDS, a dialectal phenomenon of rural-urban contrast which needs to be addressed. As we assume dialects of less populated places to converge toward regional hubs, it is indeed beneficial to test this assumption in later analyses with contemporary data if smaller communities are selected along with regional hubs. For example, Reigoldswil (BL), Maur (ZH), and Wilchingen (canton of Schaffhausen—SH) were added for this reason. Additionally, this step has led to dropping some touristic locations assumed to have changed their dialects, such as Klosters (GR), places that became suburbs, such as Pratteln (BL), and to adding more rural varieties, such as Mammern (canton of Thurgau—TG) and Linthal (GL). After consolidating external experts' opinions, the overlap of the initial set of candidate sites and the final selection was 91 out of 114.

Finally, **Figure 8** presents the conclusive 125 survey sites resulting from the synthesis of the clustering results and their sociodemographic and linguistic revision. In this figure, red sites present those selected for SDATS, while all other SDS survey sites are shown in gray. The distribution of the selected sites is more or less uniform and equidistant, similar to SDS. This means a higher density of SDATS survey sites in the alpine regions, relative to its lower population. At the same time, the alpine region exhibits a greater local variation of dialects, owing to the higher potential isolation caused by more rugged terrain. It has to be noted, however, that the qualitative requirement to have equidistant survey sites did not inform the experiment design.

5. DISCUSSION

5.1. Summary and Key Findings

Dialectometry uses clustering extensively for determining dialect areas based on linguistic similarity. However, such methods have not been utilized so far for the task of site reduction. We explored this direction with a linguist in mind who aims to revisit dialectal phenomena at representative survey sites of a previously recorded database. We propose a general pathway for incorporating a clustering procedure into the site selection methodology. Since we basically detect clusters in linguistic distance matrices and then appoint a representative survey site in (the spatial projection of) the clusters as candidates subject

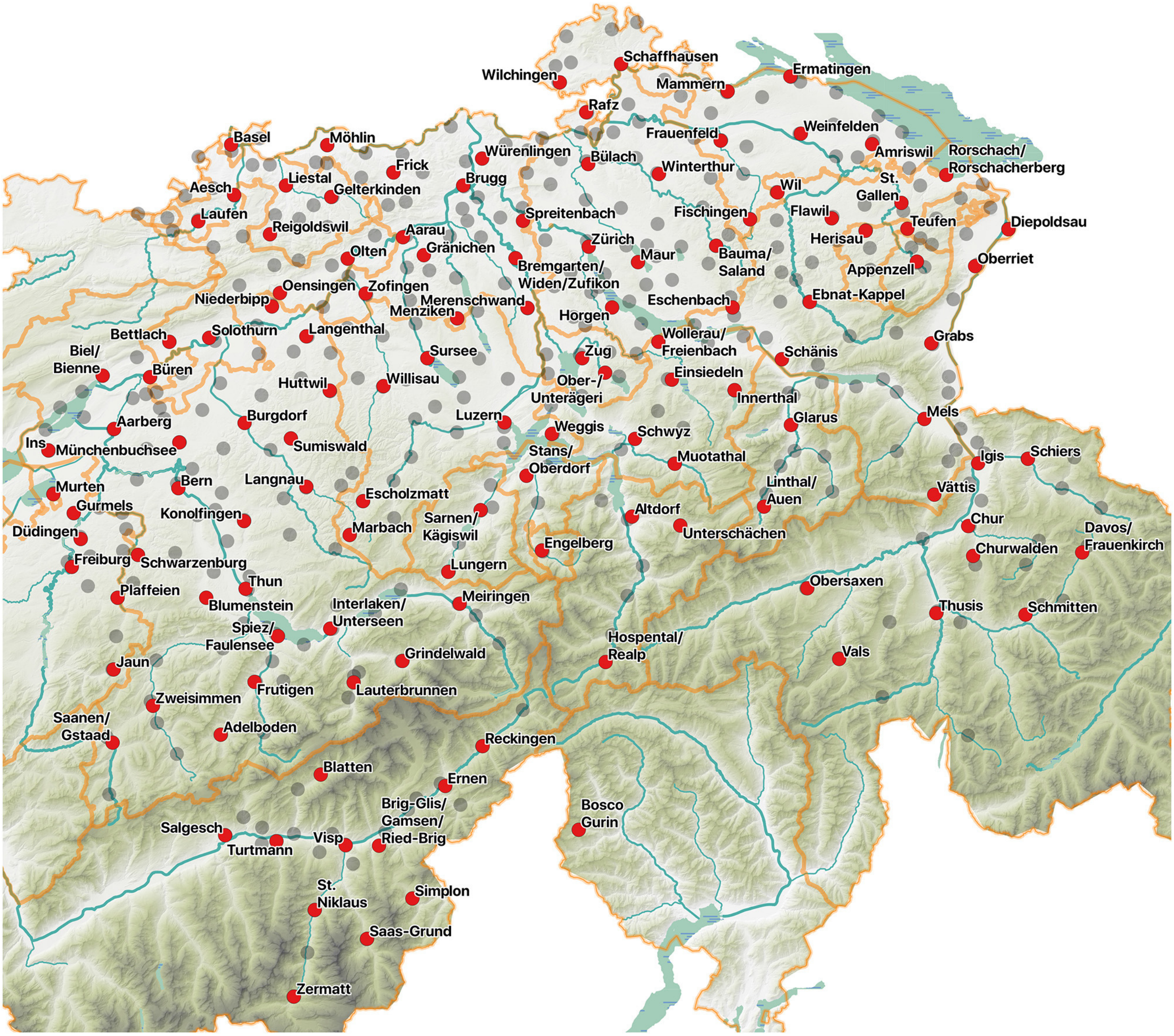


FIGURE 8 | The final selection of SDATS survey sites (in red with names), along with all SDS sites (gray). After the quantitative analysis and the qualitative revision, 125 sites are selected from the original 573.

to a qualitative revision, the methodology is appropriate for several situations. Essentially, the general methodology is based on suggestions and best practices; there is no one-size-fits-all strategy.

Rather than selecting sites based on a grid, we argue for the definition of clusters in non-spatial dimensions, where possible. We demonstrated the quantitative steps of the methodology on data from LAJ as a proof of concept and elaborated a complete application with data from SDS. These examples show that expert revision of candidate survey sites is indispensable, due to the potential bias and uncertainties in the underlying data. Due to constant language change, this expertise appreciates in value with the time elapsed since the collection of the original database.

5.2. Interpretation of the Contributions

We find that the following three intertwined aspects impact the choice of specific aspects in the methodology and play a role in the feasibility of an objective implementation:

- An optimal site reduction procedure depends on the overlap of the original and intended studies with regard to their objectives and variables.
- The dialect change that has potentially occurred between the original and intended studies needs to be considered, since the aim of most site reduction tasks is to represent the contemporary dialectal variation.
- Local representation, sought by the applied method, may crucially depend on the purpose of the intended study.

Therefore, a qualitative revision of candidate sites may overwrite previous decisions based on several considerations.

By applying the outlined methodology to two databases, we demonstrate that an arbitrary number of representative survey sites can be found within digitized linguistic survey data. The main benefit of the general methodology is to offer candidate survey sites in a quantitative framework, despite the underlying data being potentially fuzzy and uncertain. In a subsequent step, researchers are encouraged to revise the candidate survey sites according to the requirements of their intended study. Specifically, the magnitude of the potential language change that has occurred since the collection of the original data appears to impact the importance of the (partly) qualitative revision over the quantitative steps resulting in candidate survey sites.

The overlap of objectives between the SDS survey and SDATS are decreased by the sociolinguistic aims of SDATS, including SDATS' requirement of more participants per survey site, and its interest in a local colloquial dialect rather than the base dialect. Interpreting the application of the methodology to SDS, we conclude that the individual parameters of the site reduction methodology might be less important than the aims and linguistic knowledge of the researchers. Therefore, the selected 125 survey sites of SDATS are subjective to some degree.

Although not often used in dialectology, we argue that random samples are not ideal for site reduction in a project which specifically aims to capture representative linguistic variation. Even with a high number of random points, following the spatial distribution of the survey genitive sites network, random samples might not follow the distributions in linguistic variables as much as clustering solutions do by design. By selecting the linguistically central points in clusters, one can assert with a higher confidence that the selected site represents the other cluster members.

Our approach essentially also implies that, if digital data are present, it is possible to achieve the representation of the underlying data based on any number of chosen points, e.g., by taking the 20 most distinctive survey sites. This would, of course, lead to a large-scale loss of variation, and it would imply the need for an even more careful qualitative revision after partitioning the data.

5.3. Implications for Contemporary Dialectology

The methodology proposed has a number of implications for contemporary research in dialectology and beyond, for sociolinguistics, and more general language surveys. First, the automation of the site reduction process, based on the proposed methodology, allows for greater objectivity in comparison to a traditional approach where researchers have to go through the previous records or atlas data linked to the original survey sites to find the most distinct and/or representative survey sites. The availability of digitized data, clearly, opens opportunities toward faster quantitative approaches.

Second, the usage of cluster validation methods can mitigate the uncertain and fuzzy nature of the underlying data, as reflected in the clustering results. Bootstrapping and noisy clustering methods aid the estimation of this uncertainty during the clustering procedure itself, allowing researchers to adjust their site reduction methods, e.g., the intended number of survey sites, based on stability measures and the aims of their study. Most previous studies (e.g., Kelle, 2001; Christen et al., 2015)[15] used a grid approach for resampling their respective original set of survey sites and adjusted their sites manually based on expertise.

Third, partitioning clustering algorithms have specific implications. The weakness of partitioning algorithms for the classic usage in dialectometry—finding the *optimal* number of clusters—lies in their sensitivity to outliers. In this regard, however, PAM's *k*-medoid approach is more robust than the *k*-means algorithm, whereas, as seen in the application examples, UPGMA method also seems to produce unreasonable clusters. These clustering algorithms, generally considered successful in dialectology, perform differently on two data sets, suggesting that researchers have to select their methods carefully. Due to the high number of clusters in our case, however, potential outliers often become clusters on their own or together with fellow outliers, regardless of the clustering method. Nerbonne and Wieling (2018) argue for the general usage of hierarchical clustering in dialectometry based on the uncontroversial nature of dialects as hierarchically structured. We argue, however, that at local levels the original hierarchy driven by phylogenetics is not pure, and variation can be more easily overwritten by the radius of local spread of varieties increasing due to the changing contact patterns and increased mobility of the population.

A final benefit of the general methodology is its versatility. As also shown in the application to SDS, survey sites that are outliers in a linguistic sense (e.g., mountain villages in Switzerland, Norway, or Bulgaria), would be uncovered by appropriate clustering in the linguistic space, even if they are spatially embedded in an otherwise homogeneous area (e.g., the Frisian cities in The Netherlands). If the data presents a perfect continuum (e.g., parts of Sweden, as shown by Leinonen, 2010), the application of the methodology with a bootstrapping approach would result in uniformly sized clusters in the abstract linguistic dimensions and in space as well. These clusters, however, would not be very stable, as, due to the continuous nature of the data, specific boundaries between clusters would not be meaningful, and clusters in each bootstrap would be slightly different, without stable "cores." In such cases, stratifying the area with a uniform spatial grid would also be justified, and a random equidistant survey site network would necessarily represent the variation. The application of the proposed methodology would also be beneficial for smaller studies, aiming to revisit a few phenomena (or new phenomena in a similar linguistic level) in a reduced set of sites. In this case, the low number of variables intentionally overlapping with the aims of the intended study allows for a less biased cluster solution. Further, it is also appropriate to apply the methodology to data other than traditional dialect collections. Sociolinguistic studies,

[15] A similar study is currently running at the *"Deutsch in Österreich"* project (DiÖ)—https://dioe.at/projekte/task-cluster-b-variation/pp02/.

beyond obtaining survey sites in space, could successfully apply the reduction method to quantitatively appoint representative speakers within groups that are identified based on linguistic items and metadata. Moreover, the methodology may help the analysis of contemporary data, such as geotagged tweets collected and then pooled according to some criteria.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: Yves Scherrer's "Sprachatlas der deutschen Schweiz" Digitized dialect maps: http://dialektkarten.ch/mapviewer/swg/index.de.html and at the Linguistic Atlas of Japan Database: https://www.lajdb.org/lajdb_data/LAJDB_data_download001_v20180328_rev.html.

AUTHOR CONTRIBUTIONS

The idea, research questions, and conceptualization for this article was developed by PJ, AL, and CS. The data was prepared by PJ and CS. The methodology was worked out and implemented in R and QGIS by PJ. Visualization and data presentation by PJ. The interpretation of results was done by PJ, AL, and CS. The original manuscript was written by PJ, AL, and CS. All authors contributed to the article and approved the submitted version.

FUNDING

The research was funded by the Swiss National Science Foundation grant no. 181090, titled Language Variation and Change in German-speaking Switzerland: 1950 vs. 2020.

REFERENCES

Anderwald, L., and Wagner, S. (2007). "FRED–The Freiburg English dialect corpus: applying corpus-linguistic research tools to the analysis of dialect data," in *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, eds J. C. Beal, K. P. Corrigan, and H. L. Moisl (London: Palgrave MacMillan), 35–53. doi: 10.1057/9780230223936_3

Avanesov, R. I. (Ed.). (1965). *Voprosnik obščeslavjanskogo lingvističeskogo atlasa*. Moscow: Nachka.

Bailey, G. H., and Dyer, M. (1992). An approach to sampling in dialectology. *Am. Speech* 67, 3–20. doi: 10.2307/455756

BFS (2018). *Ständige und nichtständige Wohnbevölkerung nach institutionellen Gliederungen, Wohnort vor 1 Jahr, Staatsangehörigkeit (Auswahl), Geschlecht und Altersklasse*. Technical report, Bundesamt für Statistik, Neuchâtel.

Birkenes, B. M. (2019). North Frisian dialects: a quantitative investigation using a parallel corpus of translations. *Us Wurk* 68, 119–168. doi: 10.21827/5c98880d173a4

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology With R. arXiv* arXiv:1011.1669v3. doi: 10.1007/978-1-4419-7976-6

Bucheli Berger, C. (2008). "Neue Technik, alte Probleme : auf dem Weg zum Syntaktischen Atlas der Deutschen Schweiz (SADS)," in *Sprachgeographie digital–die neue Generation der Sprachatlanten. Mit 80 Karten, Germanistische Linguistik 190–191*, eds S. Elspaß and W. König (Hildesheim: Olms), 29–44.

Bucheli, C., and Glaser, E. (2002). "The syntactic atlas of Swiss German dialects: empirical and methodological problems," in *Syntactic Microvariation*, Vol. 2, eds S. Barbiers, L. Cornips, and S. van der Kleij (Amsterdam: Meertens Institute Electronic Publications in Linguistics), 41–73.

Budin, G., Elspaß, S., Lenz, A. N., Newerkla, S. M., and Ziegler, A. (2019). "The research project 'German in Austria'. Variation–contact–perception," in *Dimensionen des sprachlichen Raumes. Variation–Mehrsprachigkeit–Konzeptualisierung (Schriften zur deutschen Sprache in Österreich)*, eds L. Bülow, A. K. Fischer, and K. Herbert (Frankfurt am Main: Peter Lang Verlag), 7–35.

Burridge, J., Vaux, B., Gnacik, M., and Grudeva, Y. (2019). Statistical physics of language maps in the USA. *Phys. Rev. E* 99:032305. doi: 10.1103/PhysRevE.99.032305

Cheshire, J. A., Mateos, P., and Longley, P. A. (2011). Delineating Europe's cultural regions : population structure and surname clustering. *Hum. Biol.* 83, 573–598. doi: 10.3378/027.083.0501

Christen, H. (1998). Convergence and divergence in the Swiss German dialects. *Folia Linguist.* 32, 53–68. doi: 10.1515/flin.1998.32.1-2.53

Christen, H., Bucheli, N., Guntern, M., and Schiesser, A. (2015). "Länderen: Die Urschweiz als Sprach(wissens)raum," in *Regionale Variation des Deutschen: Projekte und Perspektiven, Chapter 25*, eds R. Kehrein, A. Lameli, and S. Rabanus (Berlin; Boston, MA: de Gruyter), 621–644.

CRAN (2020). *fpc: Flexible Procedures for Clustering. R package version 2.2-9*. CRAN, 1–164. Available online at: https://cran.r-project.org/web/packages/fpc/fpc.pdf

Cressie, N. A. C. (2015). *Statistics for Spatial Data, Revised Edn*. New York, NY: John Wiley & Sons Inc.

Cysouw, M. (2007). "New approaches to cluster analysis of typological indices," in *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, eds P. Grzybek and R. Köhler (Berlin: De Gruyter), 61–76. doi: 10.1515/9783110894219.61

Daszykowski, M., Walczak, B., and Massart, D. L. (2002). Representative subset selection. *Anal. Chim. Acta* 468, 91–103. doi: 10.1016/S0003-2670(02)00651-7

Delmelle, E. M. (2009). "Spatial sampling," in *The SAGE Handbook of Spatial Analysis, Chapter 10*, eds A. S. Fotheringham and P. A. Rogerson (London; Thousand Oaks, CA; New Delhi; Singapore: SAGE Publications Ltd.), 165–186. doi: 10.4135/9780857020130.n10

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *J. Cybernet*. 4, 95–104. doi: 10.1080/01969727408546059

Elhamifar, E., Sapiro, G., and Vidal, R. (2012). "See all by looking at a few: sparse modeling for finding representative objects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1600–1607. doi: 10.1109/CVPR.2012.6247852

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer.

Fukushima, C. (2016). "Tracing real and apparent time language," in *The Future of Dialects: Selected Papers From Methods in Dialectology XV*, eds M. H. Côté, R. Knooihuizen, and J. Nerbonne (Berlin: Language Science Press), 363–376.

Gabriel, E. (Ed.). (1985). *Vorarlberger Sprachatlas mit Einschluss des Fürstentums Liechtenstein, Westtirols und des Allgäus (VALTS): Einführung in den Vorarlberger Sprachatlas*. Bregenz: Vorarlberger Landesregierung.

Gani, W., and Limam, M. (2016). A kernel distance-based representative subset selection method. *J. Stat. Comput. Simul.* 86, 135–148. doi: 10.1080/00949655.2014.996758

Glaser, E., and Bart, G. (2015). "Dialektsyntax des Schweizerdeutschen," in *Regionale Variation des Deutschen. Projekte und Perspektiven, Chapter 4*, eds R. Kehrein, A. Lameli, and S. Rabanus (Berlin: De Gruyter), 79–105. doi: 10.1515/9783110363449-005

Goebl, H. (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Verlag der Osterreichischen Akademie der Wissenschaften.

Goebl, H. (1983). "Stammbaum" und "Welle". *Z. Sprachwiss.* 2, 3–44. doi: 10.1515/ZFSW.1983.2.1.3

Grieve, J. (2014). "A comparison of statistical methods for the aggregation of regional linguistic variation," in *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, eds B. Szmrecsanyi and B. Wälchli (Berlin; New York, NY: Walter de Gruyter), 1–34. doi: 10.1515/9783110317558.53

Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Front. Artif. Intell.* 2:11. doi: 10.3389/frai.2019.00011

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Lang. Var. Change* 23, 1–29. doi: 10.1017/S095439451100007X

Griffith, D. A. (2005). Establishing qualitative geographic sample size in the presence of spatial autocorrelation. *Ann. Assoc. Am. Geograph.* 95, 740–760. doi: 10.1111/j.1467-8306.2005.00484.x

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* (Ph.D. thesis), University of Groningen, Groningen, Netherlands.

Heeringa, W., Nerbonne, J., and Kleiweg, P. (2002). "Validating dialect comparison methods," in *Classification, Automation, and New Media. Proceedings of the 24th Conference of the Gesellschaft für Klassifikation*, eds W. Gaul and G. Ritter (Heidelberg: Springer), 445–452. doi: 10.1007/978-3-642-55991-4_48

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* 52, 258–271. doi: 10.1016/j.csda.2006.11.025

Hotzenköcherle, R., Schläpfer, R., Trüb, R., and Zinsli, P. (Eds.). (1962–2003). *Sprachatlas der deutschen Schweiz (1962–2003), 8th Edn*. Bern; Basel: Francke.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Comput. Environ. Urban Syst.* 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

Hyvönen, S., Leino, A., and Salmenkivi, M. (2007). Multivariate analysis of Finnish dialect data–an overview of lexical variation. *Liter. Linguist. Comput.* 22, 271–290. doi: 10.1093/llc/fqm009

Jaberg, K., and Jud, J. (Eds.). (1940). *Sprach- und Sachatlas Italiens und der Südschweiz*, Vol. 1–8. Zofingen: Ringier.

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Eaglewood Cliffs, NJ: Prentice Hall.

Jeszenszky, P., Hikosaka, Y., Imamura, S., and Yano, K. (2019). Japanese lexical variation explained by spatial contact patterns. *ISPRS Int. J. Geoinform.* 8:400. doi: 10.3390/ijgi8090400

Kaufman, L., and Rousseeuw, P. J. (1987). "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1–Norm and Related Methods*, ed Y. Dodge (Amsterdam: Elsevier), 405–416.

Kelle, B. (2001). Zur Typologie der Dialekte in der deutschsprachigen Schweiz: Ein dialektometrischer Versuch. *Dialectol. Geolinguist.* 2001, 9–34. doi: 10.1515/dig.2001.2001.9.9

Kellerhals, S. (2014). *Dialektometrische Analyse und Visualisierung von schweizerdeutschen Dialekten auf verschiedenen linguistischen Ebenen* (M.Sc. thesis), Universität Zürich, Zürich, Switzerland.

Knollová, I., Chytrý, M., Tichý, L., and Hájek, O. (2005). Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies. *J. Veg. Sci.* 16, 479–486. doi: 10.1111/j.1654-1103.2005.tb02388.x

Kondo, M. C., Bream, K. D., Barg, F. K., and Branas, C. C. (2014). A random spatial sampling method in a rural developing nation. *BMC Public Health* 14:338. doi: 10.1186/1471-2458-14-338

König, W. (Ed.). (2003). *SBS: Sprachatlas von Bayerisch-Schwaben*, Vol. 14. Heidelberg: Winter (Bayerischer Sprachatlas. Regionalteil 1).

Kumagai, Y. (2016). "Developing the linguistic atlas of Japan database and advancing analysis of geographical distributions of dialects," in *The Future of Dialects. Selected Papers From Methods in Dialectology XV*, eds M. H. Côté, R. Knooihuizen, and J. Nerbonne (Berlin: Language Science Press), 333–362.

Kumar, N., Liang, D., Linderman, M., and Chen, J. (2011). An optimal spatial sampling for demographic and health surveys. *SSRN Electron. J.* 1–44. doi: 10.2139/ssrn.1808947. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1808947

Kurath, H. (1949). *A Word Geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.

Lameli, A., Glaser, E., and Stöckle, P. (2020). Drawing areal information from a corpus of noisy dialect data. *J. Linguist. Geogr.* 8, 31–48. doi: 10.1017/jlg.2020.4

Lameli, A., Purschke, C., and Rabanus, S. (2015). "Digitaler Wenker-Atlas (DiWA)," in *Regionale Variation des Deutschen–Projekte und Perspektiven*, eds R. Kehrein, A. Lameli, and S. Rabanus (Berlin; Boston, MA: De Gruyter), 127–154.

Lawson, R. G., and Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *J. Chem. Inform. Comput. Sci.* 30, 36–41. doi: 10.1021/ci00065a010

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., and Messerli, J. (2020a). Linguistic fieldwork in a pandemic: supervised data collection combining smartphone recordings and videoconferencing. *Linguist. Vanguard* 6:61. doi: 10.1515/lingvan-2020-0061

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., and Messerli, J. (2020b). *Sdats*. iBros.ch. Available online at: https://play.google.com/store/apps/details?id=ch.sdats; https://apps.apple.com/ch/app/sdats/id1516597765?ign-mpt=uo%3D4

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., and Messerli, J. (2020c). *SDATS Corpus–Swiss German Dialects Across Time and Space*. Technical report, Center for the Study of Language and Society, University of Bern, Bern.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology* 74, 1659–1673. doi: 10.2307/1939924

Leinonen, T. (2010). *An acoustic analysis of vowel pronunciation in Swedish Dialects Therese Leinonen* (Ph.D. thesis), University of Groningen, Groningen, Netherlands.

Leinonen, T., Çöltekin, Ç., and Nerbonne, J. (2016). Using gabmap. *Lingua* 178, 71–83. doi: 10.1016/j.lingua.2015.02.004

Lengyel, A., Chytrý, M., and Tichý, L. (2011). Heterogeneity-constrained random resampling of phytosociological databases. *J. Veg. Sci.* 22, 175–183. doi: 10.1111/j.1654-1103.2010.01225.x

Levshina, N. (2015). *How to Do Linguistics With R: Data Exploration and Statistical Analysis*. Amsterdam; Philadelphia, PA: John Benjamins.

Linn, M. D. (1983). Informant selection in dialectology. *Am. Speech* 58, 225–243. doi: 10.2307/455229

Loos, J., Hanspach, J., von Wehrden, H., Beldean, M., Moga, C. I., and Fischer, J. (2015). Developing robust field survey protocols in landscape ecology: a case study on birds, plants and butterflies. *Biodiv. Conserv.* 24, 33–46. doi: 10.1007/s10531-014-0786-3

MacAulay, R. (2018). "Dialect sampling methods," in *The Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Hoboken, NJ: Wiley-Blackwell), 241–252. doi: 10.1002/9781118827628.ch13

Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M. (2019). *cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0*. Available online at: https://svn.r-project.org/R-packages/trunk/cluster

Maltauro, T. C., Guedes, L. P., and Uribe-Opazo, M. A. (2019). Reduction of sample size in the analysis of spatial variability of nonstationary soil chemical attributes. *Engenh. Agríc.* 39, 56–65. doi: 10.1590/1809-4430-eng.agric.v39nep56-65/2019

Manni, F., Heeringa, W., and Nerbonne, J. (2006). To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands. *Liter. Linguist. Comput.* 21, 507–528. doi: 10.1093/llc/fql040

McDavid, R. I. (1971). Planning the grid. *Am. Speech* 46, 9–26. doi: 10.2307/3087982

Meilă, M. (2007). Comparing clusterings-an information based distance. *J. Multivar. Anal.* 98, 873–895. doi: 10.1016/j.jmva.2006.11.013

Mucha, H. J., and Haimerl, E. (2005). "Automatic validation of hierarchical cluster analysis with application in dialectometry," in *Classification–The Ubiquitous Challenge. Proceedings of 28th Mtg Gesellschaft für Klassifikation, Dortmund, March 9–11, 2004*, eds C. Weihs and W. Gaul (Berlin: Springer), 513–520. doi: 10.1007/3-540-28084-7_60

Nerbonne, J., and Kleiweg, P. (2007). Toward a dialectological yardstick. *J. Quant. Linguist.* 14, 148–167. doi: 10.1080/09296170701379260

Nerbonne, J., Kleiweg, P., Heeringa, W., and Manni, F. (2008). "Projecting dialect distances to geography: bootstrap clustering vs. noisy clustering," in *Data Analysis, Machine Learning and Applications*, eds C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Berlin; Heidelberg: Springer), 647–654. doi: 10.1007/978-3-540-78246-9_76

Nerbonne, J., and Wieling, M. (2018). "Statistics for aggregate variationist analyses," in *Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watts (Boston, MA: Wiley), 400–415. doi: 10.1002/9781118827628.ch23

NLRI (1966–1974). *Linguistic Atlas of Japan (Nihon gengo chizu) (1966–1974)*. Tokyo: Printing Bureau, Ministry of Finance.

Olea, R. A. (1984). Sampling design optimization for spatial functions. *Math. Geol.* 16, 369–392. doi: 10.1007/BF01029887

Onishi, T. (Ed.). (2016). *Shin Nihon Gengo Chizu [New Linguistic Atlas of Japan: NLJ]*. Tokyo: Asakura Shoten.

Park, H. S., and Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl*. 36, 3336–3341. doi: 10.1016/j.eswa.2008.01.039

Prokić, J., and Nerbonne, J. (2008). Recognising groups among dialects. *Int. J. Hum. Arts Comput*. 1, 153–172. doi: 10.3366/E1753854809000366

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*. 66, 846–850. doi: 10.1080/01621459.1971.10482356

Ripley, B. D. (1981). *Spatial Statistics*. New York, NY: Wiley.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*. 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Scherrer, Y. (2012). *Generating Swiss German sentences from standard German: a multi-dialectal approach* (Ph.D. thesis), Université de Genève, Geneva, Switzerland.

Scherrer, Y. (2021). "dialektkarten.ch - Interactive dialect maps for German-speaking Switzerland and other European dialect areas," in *Berichte aus der digitalen Geolinguistik (II): Akten der zweiten Arbeitstagung des DFG-Langfristvorhabens VerbaAlpina und seiner Kooperationspartner am 18.06.2019*, eds T. Krefeld, S. Lücke, and C. Mutter (Munich: Korpus im Text, University of Munich).

Scherrer, Y., and Stoeckle, P. (2016). A quantitative approach to Swiss German–dialectometric analyses and comparisons of linguistic levels. *Dialectol. Geolinguist*. 24, 92–125. doi: 10.1515/dialect-2016-0006

Schmid, S., Leemann, A., Studer-Joho, D., and Kolly, M. J. (2019). Areale variation von /r/-Realisierungen in schweizerdeutschen Dialekten. Eine quantitative Untersuchung von Crowdsourcing-Daten. *Linguist. Online* 98, 11–30. doi: 10.13092/lo.98.5923

Schubert, E., and Rousseeuw, P. J. (2019). "Faster K-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms," in *Similarity Search and Applications. SISAP 2019. Lecture Notes in Computer Science*, Vol. 11807, eds G. Amato, C. Gennaro, V. Oria, and M. Radovanović (Cham: Springer International Publishing), 171–187. doi: 10.1007/978-3-030-32047-8_16

Séguy, J. (1973). *Atlas linguistique et ethnographique de la Gascogne*, Vol. 6. Paris: Centre National de la Recherche Scientifique.

Sneath, P. H. A., and Sokal, R. R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: W. H. Freeman and Co.

Sokal, R. R., and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon* 11, 33–40. doi: 10.2307/1217208

Spruit, M. R. (2006). Measuring syntactic variation in Dutch dialects. *Liter. Linguist. Comput*. 21, 493–506. doi: 10.1093/llc/fql043

Steger, H., and Schupp, V. (eds.). (1993). *Einleitung zum Südwestdeutschen Sprachatlas*. Marburg: N.G. Elwert.

Syrjänen, K. J. J., Honkola, T., Lehtinen, J., Leino, A., and Vesakoski, O. (2016). Applying population genetic approaches within languages. *Lang. Dyn. Change* 6, 235–283. doi: 10.1163/22105832-00602002

Szmrecsanyi, B. (2012). "Geography is overrated," in *Dialectological and Folk Dialectological Concepts of Space–Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, eds S. Hansen, C. Schwarz, P. Stoeckle, and T. Streck (Berlin; Boston, MA: De Gruyter), 215–231. doi: 10.1515/9783110229127.215

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr*. 46, 234–240. doi: 10.2307/143141

Trüb, R. (2003). *Sprachatlas der deutschen Schweiz. Abschlussband. Werkgeschichte, Publikationsmethode, Gesamtregister*. Tübingen: Francke.

Trudgill, P. (1974). Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Lang. Soc*. 2, 215–246. doi: 10.1017/S0047404500004358

Ueberwasser, S., and Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguist. Online* 84, 105–126. doi: 10.13092/lo.84.3849

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*. 58, 236–244. doi: 10.1080/01621459.1963.10500845

Wieling, M., and Nerbonne, J. (2015). Advances in dialectometry. *Annu. Rev. Linguist*. 1, 243–264. doi: 10.1146/annurev-linguist-030514-124930

Wilks, D. (1995). *Statistical Methods in the Atmospheric Sciences, International Geophysics, 1st Edn*, Vol. 59. Cambridge, MA: Academic Press.

Willis, D. (2020). Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: two case studies from Welsh. *Glossa* 5:103. doi: 10.5334/gjgl.1073

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 01 June 2021
doi: 10.3389/frai.2021.609970

Registerial Adaptation vs. Innovation Across Situational Contexts: 18th Century Women in Transition

Stefania Degaetano-Ortlieb[1], *Tanja Säily*[2]* *and Yuri Bizzoni*[1]

[1] *Department of Language Science and Technology, Saarland University, Saarbrücken, Germany,* [2] *Department of Languages, Faculty of Arts, University of Helsinki, Helsinki, Finland*

Endeavors to computationally model language variation and change are ever increasing. While analyses of recent diachronic trends are frequently conducted, long-term trends accounting for sociolinguistic variation are less well-studied. Our work sheds light on the temporal dynamics of language use of British 18th century women as a group in transition across two situational contexts. Our findings reveal that in formal contexts women adapt to register conventions, while in informal contexts they act as innovators of change in language use influencing others. While adopted from other disciplines, our methods inform (historical) sociolinguistic work in novel ways. These methods include diachronic periodization by Kullback-Leibler divergence to determine periods of change and relevant features of variation, and event cascades as influencer models.

Keywords: linguistic innovation, register variation, gender-specific linguistic variation, diachronic variation in language use, periods of change in language use, computational sociolinguistics, Late Modern English, historical sociolinguistics

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Justyna Robinson,
University of Sussex, United Kingdom
Mel Evans,
University of Leicester,
United Kingdom

****Correspondence:***
Tanja Säily
tanja.saily@helsinki.fi

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *24 September 2020*
Accepted: *19 April 2021*
Published: *01 June 2021*

Citation:
Degaetano-Ortlieb S, Säily T and Bizzoni Y (2021) Registerial Adaptation vs. Innovation Across Situational Contexts: 18th Century Women in Transition.
Front. Artif. Intell. 4:609970.
doi: 10.3389/frai.2021.609970

1. INTRODUCTION

The investigation of the temporal dynamics of language is becoming a growing field outside of historical linguistics. Computational modeling (such as distributional, probabilistic, neural, etc.) is adopted to trace diachronic developments. From a computational sociolinguistic perspective, our aim is to apply computational models to shed light on how sociolinguistic factors are involved in the temporal dynamics of language use. We look at Late Modern British English in the 18th century, a period of transition in social terms, considering register and gender, with a special focus on changes in women's language use. Consider, for instance, O'Brien (2009)'s examination of female writers, centered upon an analysis of the ways in which women "deployed and refashioned" (p. 2) enlightenment concepts of gender, constructing a discourse that defined and defended female intellectual and moral agency, and in the longer term enabled the development of 19th-century feminist discourse (cf. Carr, 2009, review no. 831).

As one sociolinguistic factor, *register* is known to impact language use (Biber, 1988; Halliday, 1989; Biber et al., 1999) and has been accounted for in historical linguistic analyses (Nevalainen and Raumolin-Brunberg, 2003, p. 195). Registers are defined as clusters of associated lexico-grammatical features having a greater-than-random tendency to co-occur (Halliday, 1988, p. 162) and are referred to as language use according to the situational context. In sociolinguistic terms, social interaction is realized in linguistic forms through meanings, i.e., the social context is realized by specific lexico-grammatical choices. This relation is bidirectional, i.e., a particular social context influences the lexico-grammatical choices made, while at the same time lexico-grammatical choices create a social context. However, registers are not static; in fact, studies on register formation processes have shown how registers emerge and evolve over time due to changing social contexts (Ure, 1971, 1982). The process of modernization within a society, for example, is one major trigger leading to register change and the formation of new registers (Halliday, 1988; Degaetano-Ortlieb, 2015; Teich et al., 2016). Previous research into register variation in the history of English has found that speech-based and popular written registers have been going through a gradual process of colloquialization, where they have drifted toward more oral styles, while expository "specialist" registers have developed toward the literate end of the continuum (Biber and Finegan, 1997). This research, however, has ignored other social factors (e.g., gender, social class), which alongside the situational context have a major impact on language use (Argamon et al., 2003; Nevalainen and Raumolin-Brunberg, 2003; Säily, 2016, 2018a).

Gender is one of these social factors having an impact on change in language use. As is customary in sociolinguistic research, we use the term gender rather than sex to denote this variable sociocultural construct. Previous studies have shown how the language use of women and men is distinctively involved in change, women often leading the change of the investigated linguistic phenomena (Nevalainen and Raumolin-Brunberg, 2003). Our chief interest is in the language use of women of the middle and upper classes, a social group in transition in the 18th century. According to Ylivuori (2019, p. 39, 43), the notion of gender was in flux at that time, with the early modern idea of gender as a cline between men and women being replaced by the idea of two separate genders, which encouraged heterosociability as women were thought to be naturally polite and thus to act as an improving influence upon men. This was especially the case among the upper and middle classes, whose men and women began to spend more time together in public and whose ideal of marriage also changed toward a more affectionate and informal relationship (Hay and Rogers, 1997, p. 41, 18–24). Paradoxically, women's "natural" femininity was regarded as something that required education and constant repetition in order to stick, and what exactly constituted feminine was up for debate, which gave women of the "better sort" some leeway to negotiate how they spoke and behaved, as well as opportunities to gain a better education and claim some power (Ylivuori, 2019, p. 45ff.; cf. Tieken-Boon van Ostade, 2010). We analyze the language use of these women in two different situational contexts: court trials and letter writing.

Given register and gender, we formulate the following main hypotheses:

H1 *Registerial adaptation*: due to language-external pressures in more formal contexts (court), middle and upper-class women will linguistically adapt to more formal conventions diachronically to meet social pressure (cf. Degaetano-Ortlieb, 2018)

H2 *Registerial innovation*: in less formal contexts (letters to family members), women will indicate a different linguistic behavior, perhaps even leading the change toward a more oral or involved style (cf. Säily et al., 2017b)

In our approach, we take into account the following considerations. First, similar to other studies (Gries and Hilpert, 2010), we want to broaden our understanding of the temporal dynamics in language use by considering linguistic factors as well as more than one extra-linguistic factor (here: time, register, and gender). Second, for decades in historical linguistics two things have been mainly assumed: (1) linguistic domains/levels are relatively modular and discrete, and (2) time periods are relatively fixed (cf. Nevalainen and Traugott, 2012, p. 3). These assumptions are increasingly being challenged—most prominently by those exploring the probabilistic nature of language (Bod et al., 2003; Halliday, 2004), and also due to the application of statistical methods and data mining techniques to the analysis of temporal dynamics in language (Gries and Hilpert, 2010; Degaetano-Ortlieb and Teich, 2018, 2019).

Considering the first point raised by Nevalainen and Traugott (2012, p. 3), while previous work on diachronic variation has mainly focused on one linguistic level [e.g., phonology; see also sociolinguistic (Labov, 1994, 2001) and computational sociolinguistic studies (e.g., Eisenstein, 2015; Nguyen et al., 2016)], recent studies are increasingly considering several linguistic levels and possible interplay across linguistic levels in order to obtain a more comprehensive picture of change (Bermudez-Otero and Trousdale, 2012; Broccias, 2012; Degaetano-Ortlieb and Teich, 2019; Bizzoni et al., 2020).

As for the second point, analyzing and comparing fixed time periods by pre-defining historical stages has been the standard practice (e.g., Kytö, 1993; Nevalainen and Raumolin-Brunberg, 2003; Degaetano-Ortlieb, 2015, 2018; Teich et al., 2016; Säily et al., 2017a; Degaetano-Ortlieb et al., 2019c). The rise in interest in the investigation of temporal dynamics of cultural sociolinguistic phenomena has triggered a whole wave of more exploratory, data-driven approaches targeted toward determining when particular changes occur rather than comparing predefined periods. For example, Gries and Hilpert (2008) propose a specific clustering approach to analyze the development of English targeted at single linguistic phenomena, van Hulle and Kestemont (2016) use stylometric methods to periodize literary works of Beckett, and Popescu and Strapparava (2013) characterize epochs by a statistical approach. We have designed a data-driven periodization technique based on Kullback-Leibler divergence (henceforth KLD) that allows us to detect actual periods of change from the data itself, not confined to a particular linguistic phenomenon, but across linguistic levels (Degaetano-Ortlieb and Teich, 2018, 2019). Formally, KLD measures how much two probability distributions (here: one for future and one for past language use) diverge from one another. High KLD indicates high divergence, i.e., future and past language use diverges, while low KLD indicates periods of consolidation where future and past are relatively similar to each other. Thus, peaks in KLD point us to periods of change. Moreover, interest is rising within the computational sociolinguistic community in detecting influencer (initiators of changes) and influenced (those adopting changes) groups. Recently, event cascades have shown promising results on social media interactions (Dutta et al., 2020) and conversations

(Daw et al., 2020). We adapt event cascades to model long-term diachrony.

Methodologically, we start by considering baseline models encompassing all language users of both registers (letters, court trials), comparing models of lexis, grammar, and morphology by KLD over time. We then proceed to compare gender-specific models over time. In line with H1, we assume converging trends for the more formal context (court trials), i.e., diachronically language is used more similarly across social groups, while we assume less converging trends for the informal context of family letters. To capture H2, we focus on the informal register using event cascades to investigate whether particular social groups influence others (e.g., women influencing men). Finally, we qualitatively inspect changes in the letter corpus in the broader context of the 18th century as a period of transition for women.

2. RELATED WORK

2.1. Historical Sociolinguistics

Linguistic variation and change is often socially conditioned. Sociolinguistic research has discovered, for instance, that women tend to lead language change (Tagliamonte, 2012, p. 63). Present-day sociolinguistics has typically relied on apparent-time studies of change, which make the problematic assumption that people do not change their language use as they get older. Historical sociolinguistics, spearheaded by Nevalainen and Raumolin-Brunberg (1996, 2003), has enabled the study of language change in real time in the long diachrony. It has also moderated the finding of women leading changes by pointing out that historical facts like women's lack of access to certain registers have limited their involvement in some changes, which have been led from above by men (Nevalainen and Raumolin-Brunberg, 2003, p. 131). This seems likely to also apply to stylistic change in courtroom discourse, where female attendees would have formed a small minority (Emsley et al., 2018a). Considering research on innovation and propagation, Peter Petré's pioneering work (Petré and Van de Velde, 2018; Petré and Anthonissen, 2020) shows how individual variation and diachronic change are related. Petré et al's. work combines qualitative and quantitative methods to measure the degree of grammaticalization at the level of individual attestations of particular grammatical features, also tracing lifespan change in individual authors.

While most of the research within variationist/quantitative sociolinguistics, whether present-day or historical, has focused on individual linguistic features (e.g., Tagliamonte, 2009; Nevalainen et al., 2018), there are also some large-scale studies that take a bird's-eye view of sociolinguistic variation and change in a specific corpus. These studies have typically utilized either keyword analysis or, more frequently, part-of-speech ratios (Rayson et al., 1997; Markus, 2001; Heylighen and Dewaele, 2002; Argamon et al., 2003; Newman et al., 2008; Säily et al., 2011, 2017b; Bamman et al., 2014). This work has revealed interesting and surprisingly consistent patterns of gender variation over time: whereas men's style is often characterized by an informational focus and a high frequency of e.g., nouns, determiners and numerals, women's style tends to be more oral and exhibits greater writer and addressee involvement, as evidenced by the high frequency of such features as first- and second-person pronouns, verbs, negations, and interjections (cf. Biber and Burges, 2000; Vartiainen et al., 2013). An increasing frequency of involvement features can also be seen in the colloquialization of some genres, such as personal letters, which in the fifteenth to the seventeenth centuries seems to have been led by the upper social ranks, as they "could increasingly afford to write simply to keep in touch with friends and family, for which a more oral, involved style would be in order" (Säily et al., 2017b, p. 38). The new POS-tagged version of the *Corpus of Early English Correspondence Extension*, which is equipped with social metadata, enables such research to be conducted in eighteenth-century data as well (see further 3.1.2 below).

In their studies of a number of linguistic changes in eighteenth-century English correspondence, Nevalainen et al. (2018) found that many but not all of the changes were led by women and that most of the consistently conservative individuals were men, thus supporting the sociolinguistic finding of female advantage in language change (Nevalainen, 2018, p. 257–259; Säily, 2018b, p. 242). Some of the changes they analyzed were connected to an involved style of writing, such as the incoming progressive aspect and the increase in the "embodied attribute or trait" meaning of the nominal suffixes *-ness* and *-ity*, as in *your kindness*. The latter was argued by Säily (2018a, p. 214–215) to support the claim made by McIntosh (1998, 2008, p. 231) that British culture in the later eighteenth century underwent a process of "feminization," by which McIntosh referred to an increasing concern with the feminine values of politeness and sensibility amongst those aspiring to belong to the upper echelons of society (see also Ylivuori, 2019). This could imply that middle- and upper-class men emulated the language use of the increasingly well-educated women of the same classes, who authored publications and hosted literary salons (Myers, 1990; Pohl and Schellenberg, 2003; Tieken-Boon van Ostade, 2010)—at least in some registers (cf. McIntosh, 1998). On the other hand, these women also became able to catch up with men's more nominal style and adapt it to their own purposes, which could have been reflected in the correspondence of the eighteenth-century literati (McIntosh, 1998, p. 205; Säily, 2018a). While so far considered in isolation, this study combines both views, moving toward a more comprehensive picture of the temporal dynamics involved.

2.2. Register Studies

Registers are referred to as language use according to the situational context and are defined by clusters of associated linguistic features having a greater-than-random tendency to co-occur (Halliday, 1988, p. 162). Similarly, Ferguson (1994, p. 16) states that a register is characterized by "the linguistic differences that correlate with different occasions of use." While there have been many definitions of registers with similar notions and register studies have a long tradition in linguistics (Biber, 1988; Halliday, 1988; Martin, 1992; Ferguson, 1994), "the computational study of linguistic registers was a niche area and received little attention in computational work on language overall" (Argamon, 2019). Computational work on register studies has relied on the probabilistic notion of feature co-occurrences, e.g., using classification to categorize texts according to register based on register-indicative features (e.g., Atkinson, 1992; Biber and Conrad, 2001; Argamon et al., 2008; Eisenstein et al., 2011; Teich et al., 2013, 2016), to automatically annotate register labels to account for register differences in text analysis tasks (Giesbrecht and Evert, 2009; Sharoff et al., 2010), to improve information retrieval (Morato et al., 2003; Freund et al., 2006), and for register-sensitive text generation (e.g., Reiter and Williams, 2010; Crystal, 2011; Ficler and Goldberg, 2017; Jhamtani et al., 2017).

Driven by a more theoretical perspective, there have been studies on the acquisition of registerial knowledge (Ravid and Tolchinsky, 2002; Ravid and Berman, 2009). Language users acquire knowledge on using language appropriately in particular situations by mapping relevant linguistic forms to the context of communication, considering the range of expressive options available to them. "Mastery of register appropriateness thus plays an important role in acquisition of communicative competence [...]" and well-educated individuals command a wide range of registers (Ravid and Berman, 2009, p. 2). Moreover, there have been studies on register formation and evolution using corpus-based to more exploratory data mining techniques (e.g., Ferrara et al., 1991; Biber and Finegan, 1997; Nowson et al., 2005; Herring and Paolillo, 2006; Argamon et al., 2008; Teich and Fankhauser, 2010; Teich et al., 2016; Degaetano-Ortlieb and Teich, 2019; Degaetano-Ortlieb et al., 2019a). As these studies show, register is one among many factors influencing language use, such as time, medium (written vs. spoken), and other sociolinguistic variables (e.g., age, gender), which symbiotically affect each other.

In our own previous work considering time, gender and register, we have separately shown that in formal contexts (court proceedings) middle- and upper-class women linguistically adapt to male language use at the lexical and grammatical levels (Degaetano-Ortlieb, 2018), while in less formal contexts women maintain, and possibly lead changes to, a more oral or involved style (Säily et al., 2011, 2017b; Säily, 2018a) considering the grammatical and morphological levels.

Following up on this line of research, we use exploratory data-driven methods to investigate across the linguistic levels of lexis, grammar, and morphology, our two hypotheses of gender-specific *registerial adaptation* in a formal register and gender-specific *registerial innovation* in an informal register, comparing language use of women to men across court proceedings and letters to family members.

2.3. Computational Modeling

2.3.1. Detecting Periods of Change

Borrowed from mathematics and applied in engineering fields, Kullback-Leibler divergence's popularity is growing across humanities fields as diverse as stylistics, literary studies, history, and linguistics as a measure for modeling variation. For example, Hughes et al. (2012) measure stylistic influence in the evolution of literature, Klingenstein et al. (2014) analyze language use in criminal trials, Bochkarev et al. (2014) use KLD comparing word distributions within and across languages, Pechenick et al. (2015) analyze cultural and linguistic evolution, and Fankhauser et al. (2014) demonstrate the applicability of KLD for corpus comparison at large. In our own work, we have used KLD to analyze the linguistic development of English scientific writing over time[1] (Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb and Strötgen, 2018; Degaetano-Ortlieb et al., 2019b), to investigate intra-textual variation across sections of research papers from genetics (Degaetano-Ortlieb and Teich, 2017), to analyze scientification effects in literary studies (Degaetano-Ortlieb and Piper, 2019), to detect typical features of history texts (Degaetano-Ortlieb et al., 2019c), and to investigate gender- and class-specific changes in court proceedings of the Old Bailey Court (Degaetano-Ortlieb, 2018).

With our novel method of data-driven periodization, we address a common challenge in diachronic analysis: to *determine* periods of change rather than using pre-defined periods. This is an endeavor pursued in various disciplines, such as biology, musicology, literary studies, and marketing research, as well as socio- and historical linguistics, among others. In the latter, as Nevalainen and Traugott (2012, p. 3) point out, rather than using pre-defined fixed periods to analyze linguistic diachronic change, one seeks to detect *when* changes occur on a continuous scale. In musicology, for example, to detect periods of stylistic change in popular music, Mauch et al. (2015) apply data-driven methods from bioinformatics on pre-selected audio features. In literary studies, van Hulle and Kestemont (2016) use stylometric methods with selected function words for periodization of particular prose texts. Gries and Hilpert (2008) use Variability-based Neighbor Clustering algorithms to determine periods of change for selected linguistic features [*get*-passives, verb conjugation suffixes *-(e)th* and *-(e)s*]. Ji (2010) applies Hierarchical Cluster Analysis to a corpus of Chinese focusing on selected morpho-syntactic patterns underpinning the evolution of Chinese lexis. Recently, Belinkov et al. (2019) applied periodization based on word embeddings on the Arabic portion of the OpenITI corpus[2] following Gries and Hilpert (2008)'s VNC algorithm and adapting it to Word-Embedding-based Neighbor Clustering (WENC). Most similar to our work is Barron et al. (2018)'s study of parliamentary debates on the French Revolution applying overall KLD to sequential speeches and considering how much speeches diverge over time. In our work on the linguistic development of English scientific writing, in addition to considering overall KLD tendencies, we inspect features contributing to periods of increased divergence (cf. Degaetano-Ortlieb and Teich, 2018, 2019) enabling us to analyze reasons for and effects of change.

2.3.2. Modeling Influencer Groups

Considering our second hypothesis, H2, of registerial innovation in less formal contexts, we are also interested in detecting which sociolinguistic groups might initiate a change and whether it is adopted by other groups. For this, we use the Multivariate

[1] also comparing a corpus of scientific texts [RSC (Kermes et al., 2016; Fischer et al., 2020)] to a general English corpus [CLMET (De Smet, 2006; Diller et al., 2010)] to discern change specific to scientific writing.

[2] https://alraqmiyyat.github.io/OpenITI/

Hawkes Process (Hawkes, 1971; Allan, 1976), often employed to model time-bound series, such as share trends (Hawkes, 2018) and earthquake shocks (Yuan et al., 2019), as *event cascades*, and recently also used in sociolinguistics to model turn-taking interactions in social media (Goel et al., 2016; Zhang et al., 2018; Dutta et al., 2020), as well as conversations (Daw et al., 2020). We adapt event cascades to model longer diachronic influencing trends.

3. MATERIALS AND METHODS

3.1. Data

3.1.1. The Old Bailey Corpus (OBC)

To depict the formal register of court proceedings, we use the *Old Bailey Corpus* (OBC; Huber et al., 2016) based on proceedings of the Old Bailey Court in London. These proceedings contain transcribed utterances of the court's trials spanning from 1674 to 1913. According to Emsley et al. (2018b) the City of London "required that the publisher should provide a 'true, fair and perfect narrative' of the trials" and in particular they state that "witness testimony is the most fully reported element of the trials." Thus, the witness utterances can arguably be seen as a relatively precise account of spoken English of that period (cf. Huber, 2007 on the precision of the corpus as a whole). Therefore, we opt to consider the victims' and witnesses' utterances only, excluding lawyers, judges, interpreters, and defendants. The OBC was built from a digitized version of the proceedings representing a balanced subset.

In terms of annotation, utterances and sociolinguistic information was identified semi-automatically. This procedure was quite time-consuming and definitely not a trivial task. In fact, Huber and his team developed a dedicated annotation tool which allowed them, first, to automatically detect speakers based on a list of 7,500 male and female first names (about 95% coverage), and second, to scroll through the data searching for sociobiographical information to be annotated. Witness utterances, for example, started with statements about the profession of the speakers (cf. Huber et al., 2016).

Extra-linguistic information contains speaker information including gender, age, occupation (according to the HISCO standard), social class (HISCLASS standard), speaker role (defendant, interpreter, judge, lawyer, victim, and witness), and textual information (scribe, printer, publisher). Linguistic annotation is provided at the token, lemma, and part-of-speech levels using the CLAWS7 tagset (reported accuracy of 94–95%). The version used in our studies amounts to about 14 million words and is encoded in CQP (Evert, 2005). We focus on the middle-/upper-class subcorpus according to the HISCLASS standard, with 171,084 tokens for women and 1,370,390 for men. The OBC is available through a CQPweb platform.[3]

3.1.2. Tagged Corpus of Early English Correspondence Extension (TCEECE)

The *Corpora of Early English Correspondence* (CEEC; Nevalainen et al., 1998–2006) were compiled to facilitate research in historical sociolinguistics. The genre of personal letters was chosen by the compilers for two reasons. Firstly, letters are a "speech-like" genre (Culpeper and Kytö, 2010, p. 17) resembling spoken conversation, which is the primary medium of social interaction and hence of interest to sociolinguists, who see it as the hotbed of change. Secondly, correspondence is a genre that would have been available to anyone who was literate, which means that by focusing on letters it was possible to achieve a wider social representativeness than by selecting texts written for publication, which would mostly have been authored by highly educated men. Nevertheless, a bias toward these men is evident even in the CEEC: they were the most literate social group, and their letters were considered important enough to be preserved and later edited. The corpora are based on published original-spelling editions of letters, which were sampled and digitized by the corpus team. This approach enabled the collection of millions of words of text but has the drawback of copyright issues, due to which only part of the corpora have been published. The extra-linguistic information on the social background of the informants, compiled by the team based on the editions as well as other historical and biographical sources, includes, e.g., gender, social rank, domicile, and the relationship between the writer and recipient of the letter (Raumolin-Brunberg and Nevalainen, 2007). About a quarter of the informants are women.

While the original corpus covered the period of c.1410–1681, its eighteenth-century *Extension* (CEECE) extends the end date until 1800. In the present study, we use the POS-tagged version of the CEECE, or TCEECE, which comprises about 2.2 million words in 4,923 letters sent by more than 300 individual writers. Prior to POS-tagging it with CLAWS, the spelling of the corpus was normalized using VARD (Baron, 2011a,b) along with some additional manual normalization, including abbreviations; however, as the normalization only targeted sufficiently frequent items, some orthographic variation still remains. The accuracy of the POS-tagging with the CLAWS5 tagset is c. 94.7% (Saario and Säily, 2020). We use a subcorpus of the TCEECE consisting of letters written between nuclear family members, which provides an interesting counterpoint to the formal speech-based register represented by the OBC. To match the OBC data, we further narrow down the corpus by focusing on men and women of the middling and upper social ranks during the time period 1720–1799. The size of this subcorpus is about 500,000 words, of which women's letters comprise 38.8% (cf. Kaislaniemi, 2018, p. 56).

3.2. Modeling Variation

3.2.1. Measuring Divergence Between Language Uses

Recently, research in linguistic variation and change has increasingly relied on information-theoretic approaches. In particular, relative entropy formalized as Kullback-Leibler divergence (Kullback and Leibler, 1951) has proven effective to measure divergence between two probability distributions, A and B, derived from linguistic feature sets (cf. Fankhauser et al., 2014 using words). In our case, we consider three types of linguistic feature sets: lexical (word), syntactic (part-of-speech tag), and morphological (suffix). Given these levels, we define a feature set viewing a corpus as being realized as a probability distribution at

[3] Landing page: OBC (V2.0) https://www.clarin.eu/showcase/old-bailey-corpus-20-1720-1913.

one of these linguistic levels. Basically, KLD measures the number of additional bits needed to encode a given distribution A with another distribution B given a set of features (see Equation 1). Note that the set of features used with KLD can be quite vast.

$$D(A||B) = \sum_i p(feature_i|A)log_2\frac{p(feature_i|A)}{p(feature_i|B)} \qquad (1)$$

The probability of the ith linguistic feature (e.g., a word or suffix) in A, $p(feature_i|A)$, and the ith feature's probability in B, $p(feature_i|B)$, are used to measure the amount of additional bits needed. The sum over all features gives an overall divergence measure, namely KLD $D(A||B)$, which is always positive. The higher the KLD for A given B, the more the two distributions diverge. In addition, Jelinek-Mercer smoothing is used (lambda at 0.05; cf. Zhai and Lafferty, 2004; Fankhauser et al., 2014) to assign a non-zero probability to unseen features and improve the accuracy of feature probability estimation. In Degaetano-Ortlieb and Teich (2019, section 3.2.1), we give a detailed explanatory description of KLD based on a concrete calculation example.

Moreover, KLD is an asymmetric measure. Thus, the directionality of a comparison matters, i.e., A given B might result in a different value than B given A. This is particularly useful when considering language use. For example, a layperson might well be understood by an expert (e.g., in patient-doctor conversations), $D(patient||doctor)$, while an expert's language use might be difficult for a layperson to understand, if the expert uses his/her usual field-specific language (e.g., a doctor using specialized medical terminology), $D(doctor||patient)$.

Besides an overall indication of divergence between A and B, we can also inspect the individual feature weights, calculated by pointwise KLD (see Equation 2).[4] This allows us to inspect which features are primarily associated with a divergence, i.e., those features needing a (relatively) high amount of additional bits for encoding, and thus, strongly contributing to variation between A and B.

$$D_f(A||B) = p(feature_i|A)log_2\frac{p(feature_i|A)}{p(feature_i|B)} \qquad (2)$$

The feature weights obtained from KLD can be directly interpreted as bits of information. The more bits a feature needs to be encoded, the more typical it is for A in comparison to B and can thus be determined to be a relevant feature of variation for A when compared to B.

3.2.2. Data-Driven Periodization

Based on KLD, we have developed a novel data-driven periodization to determine periods of change. The approach has the following components: (1) comparison of adjacent years by KLD for the linguistic levels selected, (2) relatively unconstrained feature selection across linguistic levels, and (3) inspection of features involved in change with high contribution to the overall divergence (KLD).

For feature selection, we opt to have a relatively unconstrained selection, i.e., rather than preselecting linguistic features known to be possibly involved in change, we use ngram sequences. In particular, we consider the lexical level by selecting all words (unigrams), the grammatical level by selecting pos-trigram sequences, and the morphological level based on a list of suffixes. The choice for trigrams was made after experimenting with different ngram sizes: bigrams proved to be too short to depict phrase/clause structure, and fourgrams and fivegrams are too long, leading to sparse data. In fact, pos-trigrams have also proved to work well in other diachronic studies (Culpeper and Kytö, 2010; Kopaczyk, 2013; Degaetano-Ortlieb and Teich, 2016; Degaetano-Ortlieb et al., 2019a). For suffix selection and extraction, we rely on experts' linguistic knowledge. In total, we consider 30 suffixes[5] selected by manually revising extracted lists from the corpora to ensure data quality. Note that any kind of linguistic unit could be used for comparison (e.g., phoneme, morpheme, word, etc.).

Comparison of adjacent years by KLD is illustrated in **Figure 1**. Basically, we slide over the timeline, comparing a range of years preceding and following a selected year with KLD. This allows us to find peaks and troughs in KLD which indicate a change. The procedure is operationalized as follows (also illustrated in **Figure 1**):

(1) Select a year i (or range of years, if the publication is not yearly) as a gap and a window size n of preceding (PAST: $i-1, ..., i-n$) and following (FUTURE: $i+1, ..., i+n$) years (e.g., 20 years);
(2) Calculate KLD for the PAST and FUTURE in both directions, i.e., divergence for a language model of the PAST given a language model of the FUTURE, $D(PAST||FUTURE)$, and divergence for FUTURE given PAST, $D(FUTURE||PAST)$;
(3) Slide to the next year and repeat (2).

In Degaetano-Ortlieb and Teich (2018), we experiment with different window sizes, showing how more fine-grained selections help to detect more subtle changes (e.g., a window size of 5 years), while coarser selections (e.g., a window size of 20 years, chosen here as it is assumed to cover a generation) lead one to inspect more general trends. Considering directionality, modeling PAST given FUTURE allows us to inspect outdated language use, and modeling FUTURE given PAST, more innovative language use.

Iterating over the years, we obtain a curve of KLD values showing a trend line for past language use as well as future language use. Peaks in KLD indicate periods of change; troughs point to periods of consolidation where the past and future are more similar to each other. This allows us to inspect at which particular point in time changes occur.

In addition to these overall trends, investigating individual feature contribution allows us to gain more profound insights into the kinds of change in the indicated periods. As we are dealing here with multiple bi-class comparisons, i.e., for one direction (e.g., FUTURE given PAST) at each gap one comparison

[4] Note that pointwise KLD can also result in negative values, i.e., features have a negative contribution for A.

[5] *-able, -age, -al, -ance, -ant, -ary, -ate, -eer, -ence, -ent, -er, -ful, -hood, -ian, -ible, -ic, -ion, -ism, -ist, -ity, -ive, -ize, -less, -ment, -ness, -ology, -ous, -ship, -tude, -ure.*

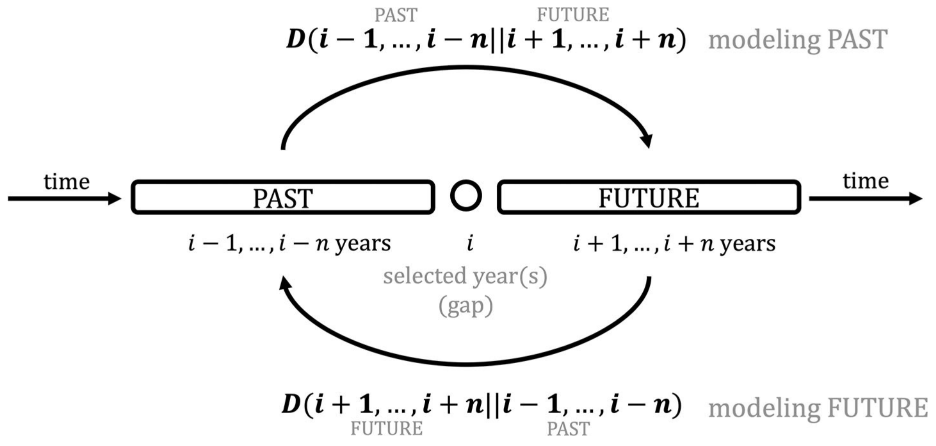


FIGURE 1 | Data-driven periodization with Kullback-Leibler Divergence.

of 20-year windows across ∼40–60 years, one has to carefully choose how to inspect the many feature rankings in a meaningful way. One option is to inspect which features show high variation in their contribution, e.g., words having a high contribution to KLD only at particular points in time. For this, a standard deviation calculated across the feature rankings can be used. Another way of inspecting relevant features across comparisons is by ranking based on the feature weights of one particular year, allowing us to inspect more confined year-specific trends.

3.2.3. Detecting Influencer Groups by Event Cascades

Event cascades are a series of events marked on a temporal axis and having some form of self-exciting pattern (see **Figure 2**). In simpler terms, they are sequences of events happening with some form of domino effect. For example, the first shock of an earthquake happens randomly, while the succession of after-shocks happens only as a consequence of the first one, i.e., the first shock initiates a cascade of events. In conversations, one influential actor can start a change (a topic, a grammatical pattern), which others take up and re-use. The early and later adoptions of a new term on social media are a typical example of such kinds of cascades (cf. Eisenstein, 2019).

Since we are essentially trying to understand who influences whom in a social network, the goal of our model is to estimate the parameters $\alpha_{i\to j}$ for all *i*,*j* in the population (Linderman and Adams, 2014). The excitation parameters can thus be represented as a matrix with $\alpha_{i\to j}$, showing the excitation intensity of events from source *i* to target *j* (see Equation 3). In turn, each node in the network is taken as a source and target of the others:

$$\lambda_t^{(j)} = \lambda_0^{(j)} + \sum_{e\,:\,t'_e<t} \alpha_{s_e\to j}\kappa_{s_e\to j}(t-t'_e), \qquad (3)$$

where $\lambda_t^{(j)}$ is the excitement intensity function on the node *j* from all other nodes at time *t* (how much other nodes have influence on *j* at time *t*), α is a scalar excitement parameter, s_e indicates the source of event *e*, and $\kappa_{s_e\to j}$ is a kernel decay function monotonically decreasing through time, constrained to integrate to 1 over positive arguments—the further away in time, the weaker the influence is expected to be.

The fundamental idea is that in a multi-party dialog, some speakers have a higher degree of influence on the style of the others: they start "event cascades" in the conversation. If actor *X* in a conversation repeatedly starts a topic or particular linguistic use, subsequently used by *Y* and *Z*, we may assume an influence from *X* to *Y* and *Z*—although this is not necessarily the case.[6] The final result is a matrix of influencers and influenced. In our case, we model female and male language use of different time periods as different actors, i.e., the data is modeled as a series of concatenated pair-wise interactions between women and men sequentially within the same period. We divide periods into 20 years to match our KLD analysis. For example, women in the 1740–1759 period are modeled with men of the same period.

The event cascade's goal is to measure the intensity of the influence of *i* on *j* for a specific time interval Δt and is modeled as a sum over *B* simple basis models (Equation 4), as the ones in **Figure 2**:

$$\kappa_{i\to j}(\Delta t) = \sum_{b=1}^{B} g_b^{(i\to j)}\phi_b(\Delta t). \qquad (4)$$

where $\phi_b(\Delta t)$ is the basis model (an impulse function that sums to one) and g_b are the dyad-specific weights over the basis models.[7] This allows us to see whether one sociolinguistic group in the exchange has a particular influence over the others. The cascades these exchanges produce look like the

[6] It is important to stress here that the concept of influence is to be held as likely, and not absolutely ascertained: we are de facto observing correlation, not causation. For example, it may be that all parties are being influenced by the same hidden source with a delay, i.e., early adopters would not be influencing late adopters. When we say that *X* influences *Y*, we mean that their behavior is consistent with that of an event cascade initiated by *X* and continued by *Y*.

[7] In this case we will not directly consider one-to-many influences.

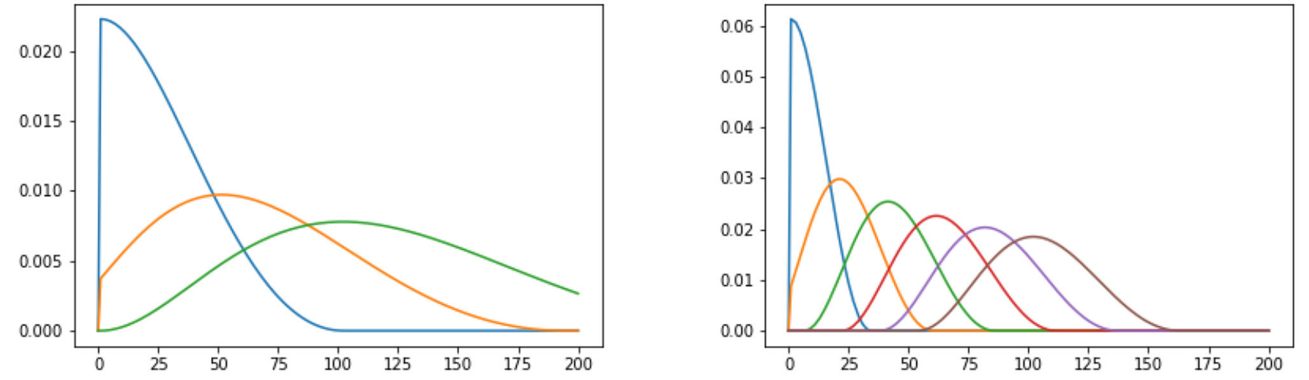


FIGURE 2 | An abstract self-exciting cascade with 3 (left) and 6 (right) bases, or impulses. Impulse duration on the x-axis and intensity on the y-axis. Each impulse influences the intensity and duration of the next one. More impulses also mean faster and more intense impulses: the process is self-exciting.

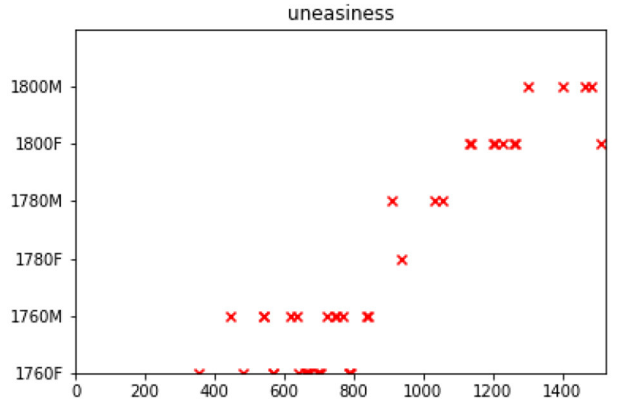


FIGURE 3 | An event cascade for the word *uneasiness*. Sociolinguistic groups on the y-axis, chronologically ordered series of letters on the x-axis.

ones in **Figure 3**, showing how words (such as *uneasiness*) are introduced in earlier periods and then continue to be used later on. Here, we can observe women starting the trend (1760F), influencing also both men and women of later periods. To obtain an overall impression of influencer and influenced groups, we use a heatmap visualization based on Equation (4) which shows the intensity of influence across groups.

4. INVESTIGATING GENDER-SPECIFIC REGISTERIAL ADAPTATION AND INNOVATION

4.1. Diachronic Tendencies Across Registers and Linguistic Levels

Using diachronic periodization with KLD, we start comparing diachronic baseline models of both the court and letter corpora (focusing on the middle and upper classes) across lexis, grammar, and morphology.

Figure 4 shows converging trends over time for both corpora. For lexis (**Figure 4A**), the order of magnitude of the decrease varies across registers, the court proceedings showing lower divergences compared to the letters, indicating a more consolidated vocabulary in court in comparison to a more varied vocabulary in the letters, which, however, might also be an effect of spelling variation still present in the letter corpus for less frequent lexemes.[8] At the grammatical level (**Figure 4B**), both registers show a similar decreasing trend, with convergence around the mid-1750s. At the morphological level (**Figure 4C**), the strong decrease in the 1740s for the court corpus is related to data sparsity in the preceding years used for modeling, but basically for both corpora the trend is a converging one.

The converging trend across linguistic levels is in line with previous work on the evolution of genres, registers and styles in English. For example, Claridge (2012, p. 82, 90) has shown trends toward standardization and regularization during 18th century English, arguing that enforcing and maintaining standardization is one of the functions of written, published language and that these written usages could then also promote the standardization of the spoken language (Milroy and Milroy, 1991, p. 35, 60, 64). Here, we see similar trends for a formal spoken register (court proceedings) and an informal written register (letters to family members). Considering a communicative perspective, there is evidence that these converging tendencies across registers are related to the fact that language users strive for efficient communication. Convergence is one effect leading to achieve this goal. Consider results from Degaetano-Ortlieb and Teich (2019) and Bizzoni et al. (2020), who show a converging tendency in scientific writing for 17th–18th century English at both the lexical and the grammatical level. They argue that a decrease in

[8] Approximately two thirds of the distinct word forms or types (typically the least frequent forms) cannot be automatically mapped to the *Oxford English Dictionary*, even if we exclude words tagged as foreign or proper nouns (Säily and Mäkelä, 2019). While we do not have similar information for the OBC, it is to be expected that the proportion would be lower there.

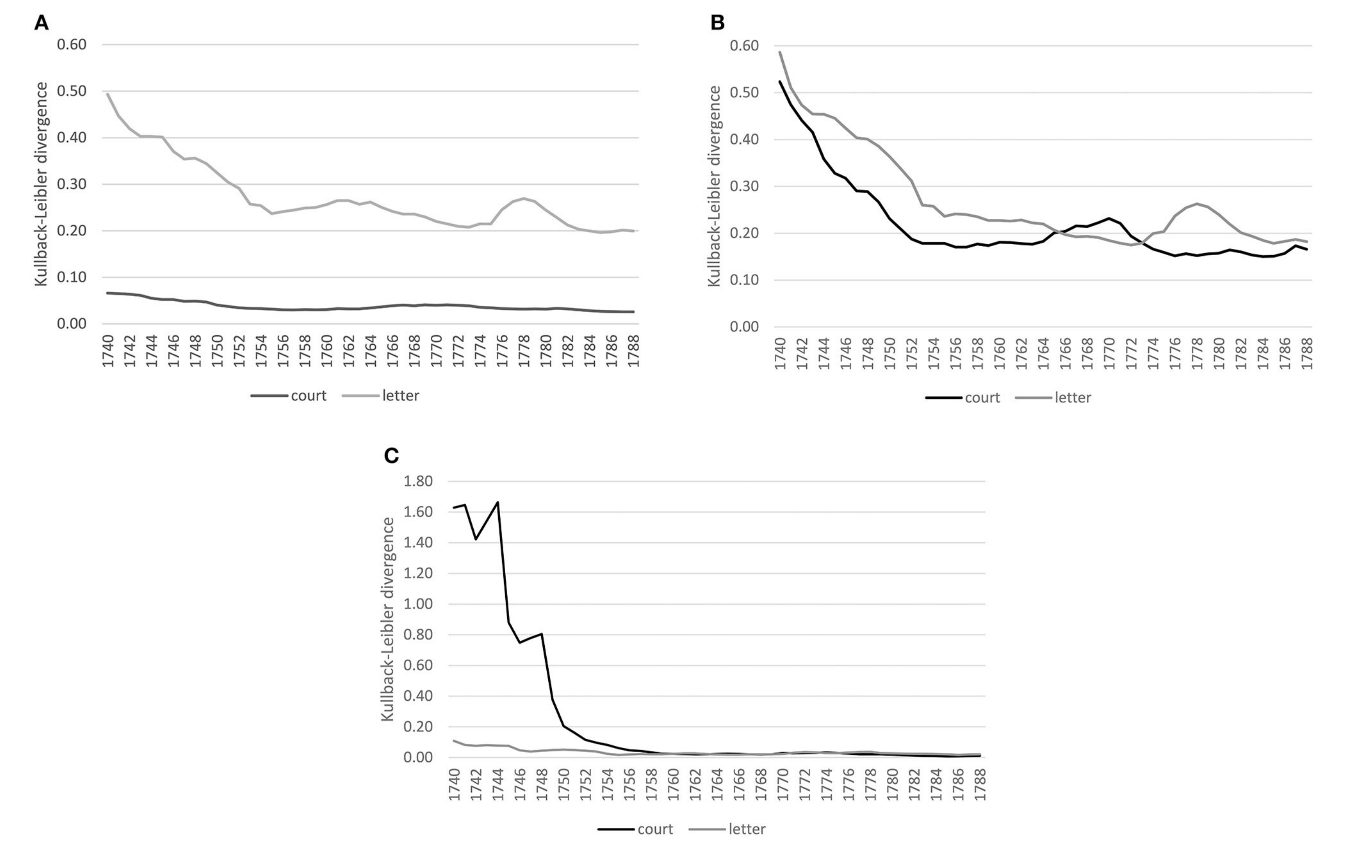


FIGURE 4 | Baseline models for the court and letter corpora across linguistic levels. **(A)** Lexis (word-unigram). **(B)** Grammar (part-of-speech trigrams). **(C)** Morphology (suffixes).

variation, i.e., convergence on particular options, is beneficial for communication. The entropy, i.e., the uncertainty about which linguistic items to use (in terms of production) or expect (in comprehension), is reduced and shared conventionalized options arise over time among language users. Despite change in language use related to converging trends leading to conventionalization, change is clearly also brought about by innovations. However, De Smet (2016) shows how conventionalization is a precondition of innovation. One aspect that is still understudied is how innovations come about and who leads those changes.

Taking up a sociolinguistic perspective considering gender, we further investigate how gender-specific groups might change their language use over time across registers, something not captured by the baseline models as they comprise all language users. We focus on middle- and upper-class women and men.

4.2. Gender-Specific Diachronic Tendencies

Here, we investigate our two hypotheses of gender-specific *registerial adaptation* in the formal contexts of court proceedings (H1), and *registerial innovation* in the informal setting of letters to family members (H2), both across lexis, grammar, and morphology. Diachronic periodization is used to determine periods of change and derive relevant features of variation. Event cascade models are applied to determine gender-specific influencer groups. Detailed micro-analytical sociolinguistic inspections are presented to elaborate on the computational findings.

4.2.1. Lexis

First, we inspect how language use of the future has changed from past language use for both the court and letter corpora at the lexical level. For this, we select 1 year, and compare the preceding 20 with the following 20 years using KLD. This is done for all years, sliding over the timeline toward the future (see section 3.2.2). **Figures 5A,B** show how male language use converges over time in both registers. Female language use seems to converge at first, until the mid-1750s/60s, with an increasing diverging tendency afterwards. Thus, women change their language use over time, while male language use increasingly converges—a tendency that applies to both registers.

Second, we ask how different the language use of women vs. men is for the same period of time (contemporary models). For this, we compare across gender the same 20 years by KLD, sliding again over the timeline toward the future. **Figure 5C** shows that in the court setting, contemporary female and male language use (see the light gray line) diverges increasingly until the mid-1750s and stabilizes afterwards with small ups and downs. Comparing

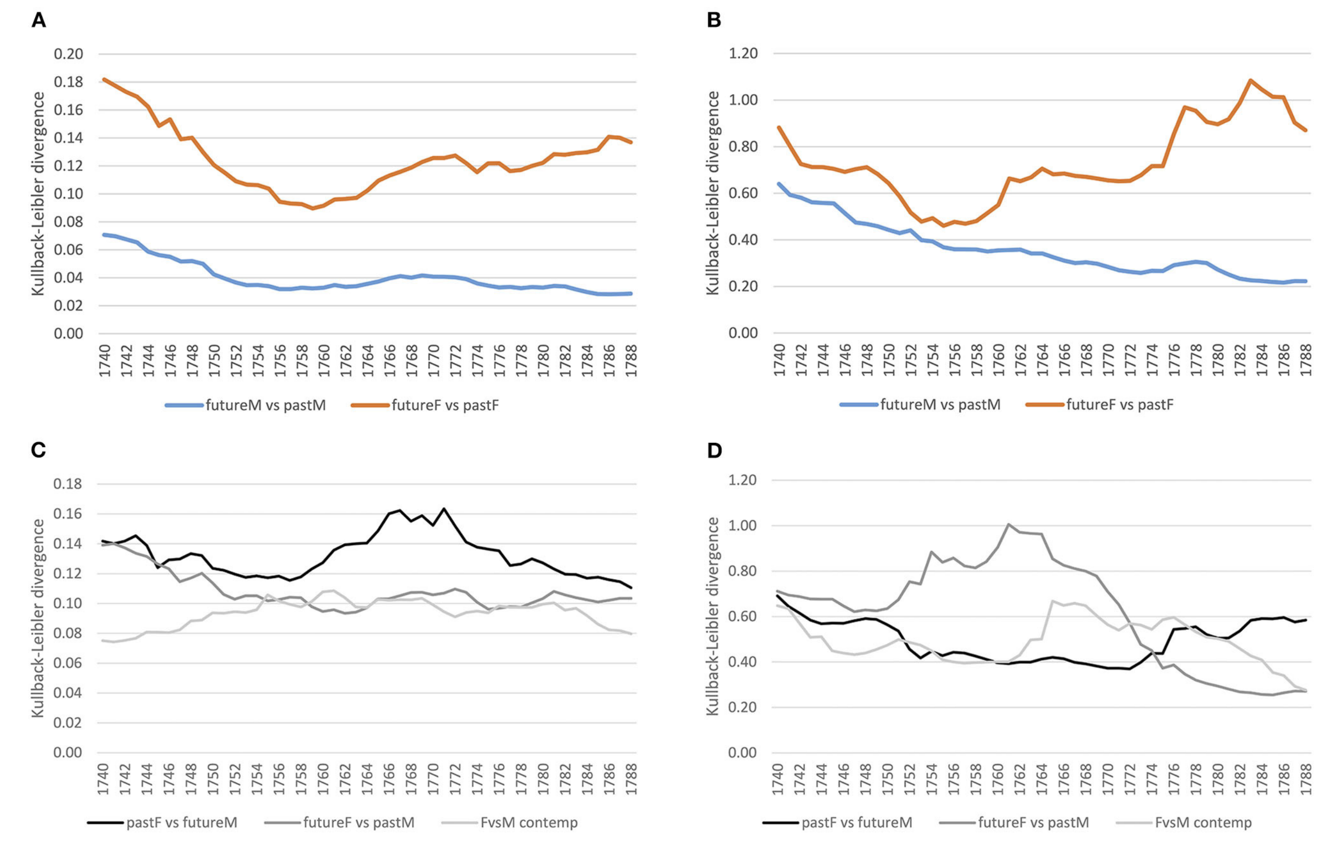


FIGURE 5 | Gender distinguished, diachronic and contemporary models (court vs. letter) at the lexical level (word-unigram). **(A)** Court: within gender comparison. **(B)** Letter: within gender comparison. **(C)** Court: diachronic and contemporary comparison. **(D)** Letter: diachronic and contemporary comparison.

this to the letter corpus (see **Figure 5D**, light gray line), the contemporary model converges at first until the end of the 1750s, diverges in the 1760s/70s, and converges further in the 1780s. In both registers, this seems to reflect ongoing change at the lexical level between male and female language use.

Third, we can inspect how female language use diverges from or converges to male language use of the future and past. This is an important perspective, because change is ongoing, so a contemporary model will miss possible adaptation tendencies. For example, if future female language use converges with past male language use, then women might have adapted to language use of males in court after having been possibly exposed to that language for a while. If this is the case, and as we have seen that future and past male language use converges over time (see **Figure 5A**), at some point past women and future men should converge as well. While social pressure of conforming to particular conventions might arise in the court setting, in the informal setting of letters to family members we assume a less stable converging trend.

Comparing both registers, we clearly see registerial differences. First, considering past female vs. future male language use in the formal court setting (see black line in **Figure 5C**), there is a rise in divergence around the 1760s,[9] while in the informal letter setting (see **Figure 5D**), divergence is relatively stable. Second, considering future female vs. past male language use (dark gray line), in the court corpus divergence decreases continuously, stabilizing around the 1750s, while in the letter corpus there is a peak in divergence from the 1750s to the mid-1760s with a steep decrease afterwards. In conclusion, in court we see registerial adaptation as women change their language use converging to men,[10] while in the letter corpus change points to registerial innovation, future women diverging from the past. Note that innovation takes place within a particular time period (the peak in **Figure 5D**), after which women and men converge again. Men, instead, seem to be more conservative (relatively stable divergence of pastF vs. futureM).

To inspect whether, during the innovative period, one gender's language use has an influence on the other's, we use event

[9] That is, 20 years before 1760 diverge from 20 years after 1760 by 1.0 bits, 1/3 more than 20 years before and after 1740.

[10] Past female language use differs from future language use of both men and women, cf. rising tendency of futureF vs. pastF in **Figure 5A**; future female language use converges to male.

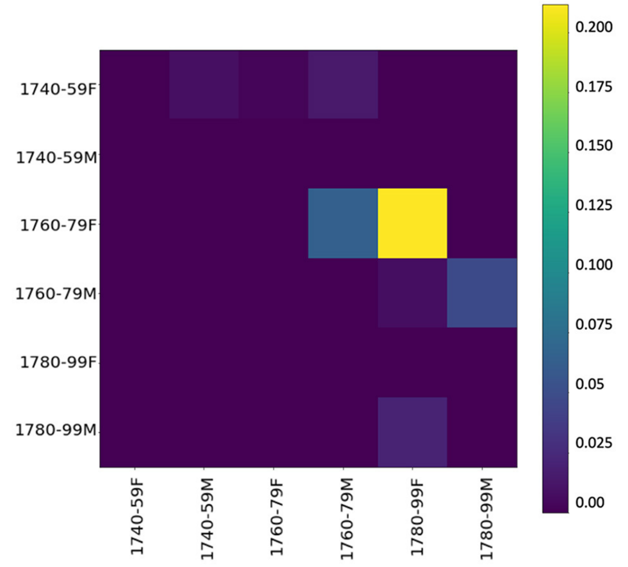


FIGURE 6 | Influencer heatmap of event cascades over time at the lexical level (500 most frequent words).

cascades (see section 3.2.3),[11] considering far-reaching influences (over the whole dataset), which allow us to see that influences tend to cascade down to the next period. **Figure 6** shows the influencing gender-and-period groups on the y-axis and the influenced groups on the x-axis—the higher the value, the stronger the influence. Basically, if a group (e.g., females 1740–1759) starts several cascades, it has a relevant influence. The yellow square in the heatmap shows a strong chronological influence by women of the 1760/70s on women of the following period. A less strong but still visible influence is on men of the same period as well as the following period, confirming our assumption derived from KLD tendencies in **Figure 5D**. Especially in the central period (ranging from 1760 to 1779), women appear to have a tendency to start using new words, both in the functional and in the content realm. Thus, it is mainly female language use at the lexical level (including both content and function words) in the informal setting of letters to family members that influences male language use.

4.2.1.1. Micro-Analysis: Lexis

As we have seen distinct language use of women and men in the period of the 1760s and 1770s (see peak in the contemporary model in **Figure 5D**, and heatmap of event cascade in **Figure 6**), we look at the features that contribute to an increase in KLD in the contemporary model. Comparing **Figures 7A,B**, we clearly see a more involved personal style for women than for men. Women distinctively use in comparison to men the personal pronouns *I* and *you*, negation (*not*, *never*) and conjunctions of contrast and concession (*though*, *but*) as well as mental verbs (*think*, *wonder*); see example (1). These features also indicate a more verbal style of writing. Men, instead, distinctively use the determiner *the*, prepositions (*upon*, *at*), and the relativizer *which* pointing to nominal style as well as first-person plural pronouns (*we*, *our*, *us*); (2). Whereas (1), written by a wife to her husband, clearly concentrates on interpersonal relationships using affective language and ego- and addressee-involvement, example (2) from a husband's letter focuses on narrating what he has done or observed with a third party, using exclusive *we*, which cannot in this context be regarded as an involvement feature. Thus, middle- and upper-class family letters in this period exhibit the oft-observed distinction between personally involved women and informatively oriented men, while the register as a whole moves in the involved direction led by women (cf. the "feminization" of McIntosh, 1998), shown also in the cascade analysis.

(1) **I** declare **I** should be rejoiced, was there no occasion, to write on things of more consequence, **as I never** wish to give **you** vexation, however my Duty to your Son obliges me to speak sometimes of things **I** know **you** don't like to hear and yet in fact your **own** interest is concern'd **as** much **as** his, **I** mean in regard to the payment of your Sisters Fortune—**I never think** of it **but** it leaves a dead weight on my **Heart**, and **I** cant help saying that it is a most cruel thing in **you** to keep runing up the interest **as you** do […]

(Eliza Taylor née Pierce to her husband, Thomas, January 29, 1766; PIERCE_028)[12]

(2) Palmer & I took **our** horses on Friday & rode to **the** Town of Dock, 2 miles, & to **the** passage **which** I have marked. From thence **we** saild to Lord Edgcombes gates & walkd over a fine lawn to **the** house, **which** is about halfway up **the** hill. **The** stone of this country is too hard & rough to work to a truth, as **we** masons say. Its colour too, **which** is a reddish black. being all really marble mix'd with a very white lime, is not agreeable to **the** eye, & **the** house being old, with 4 octagons newly added to **the** angles, makes a better appearance **at** a distance than near.

(Roger Newdigate to his wife, Sophia, October 17, 1762; NEWDIGA_037)

The female subcorpus in the 1760s–70s includes both lesser-known professionals and gentry like Mrs. Taylor in (1) and literary figures like Lady Mary Wortley Montagu, and their language use is in part surprisingly similar. McIntosh (1998, p. 205) argues that at the same time as more and more women became published authors, the social roles of women within the family became more restricted as the idealized "sentimental family" locked women inside the private sphere. We can see that Mrs. Taylor writes quite deferentially and considerately to

[11] Over the most frequently used words in each period—thus including a good deal of stylistically informative function words. We also tried this with more low-frequency content words, obtaining essentially similar results in terms of overall influence.

[12] Examples are given in their original spelling. Boldface has been added, while italics (if any) are as printed in the letter edition and probably reflect underlining in the original manuscript. PIERCE_028 is the unique identifier of the letter in the corpus.

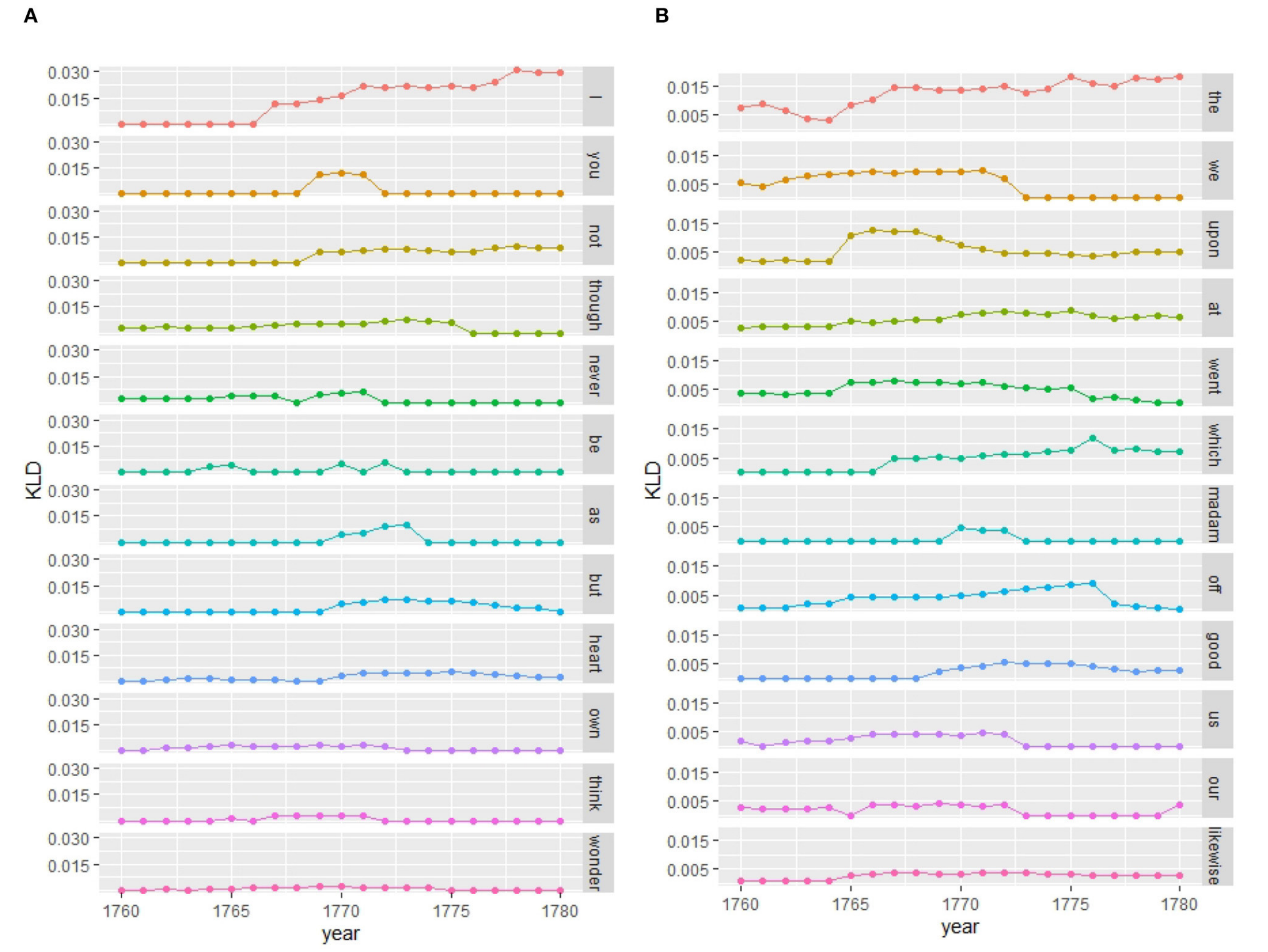


FIGURE 7 | Features contributing to period of change (from top to bottom showing highest to lowest contribution)—contemporary model women vs. men (word-unigram). **(A)** Women. **(B)** Men.

her husband, while also committing the face-threatening act of trying to tell him what to do in a financial matter, over which she as a woman has no legal control. This is one of the contexts that seem to intensify the use of involvement features in family letters of the time, with women bringing up important issues while presenting themselves as loving wife-mothers driven by their feelings (cf. Ylivuori, 2019, p. 77–82). The feminized ideal of the sentimental family also influenced male writing, and even Sir Roger of example (2) has some affective and interpersonally involved passages in his letters to “My Dearest Sophy.”

4.2.2. Grammar

Taking up the same approach for grammar as for the lexical level, we first compare female and male language use separately, considering future compared to past language use. In the formal court setting, we clearly see a converging tendency for both women and men, stabilizing around the 1750s (see **Figure 8A**). For the letter corpus, however, while male language use converges over time, women seem to converge at first until the mid-1750s, but increasingly diverge in their use of grammar afterwards compared to previous years (see **Figure 8B**). Thus, while men converge in the use of grammatical patterns in both settings, women change their use of grammar around the 1770s in the informal setting.

Considering contemporary language use of women and men in the court corpus, it is quite stable (see **Figure 8C**). The diachronic models of pastF vs. futureM and futureF vs. pastM confirm a relatively stable use of grammar in the formal setting. In the letter corpus, on the other hand, there is a period of change in the use of grammatical patterns around the mid-1760s/70s, where the language use of women and men differs as depicted by the peak in divergence. This period of change is also shown in the diachronic models: female language use before that period diverges from that of past men and past female language use diverges from that of future men after the period of change.

Looking at the event cascades for pos-trigrams (see **Figure 9**), again we can confirm an influencing trend of women toward men, especially in the period between 1760 and 1779. These results essentially mirror the findings at the lexical level, showing

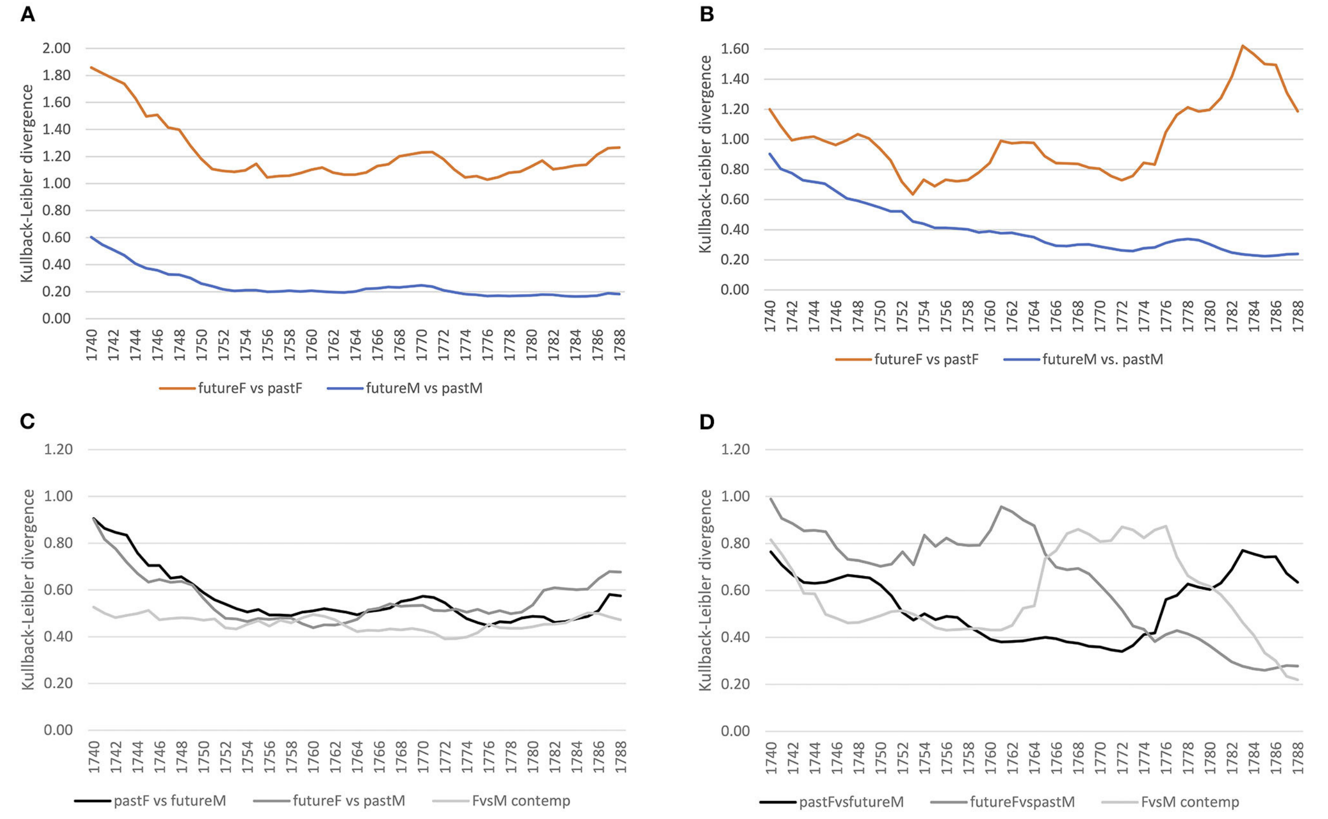


FIGURE 8 | Gender distinguished, diachronic and contemporary models (court vs. letter) at the grammatical level (pos-trigrams). **(A)** Court: within gender comparison. **(B)** Letter: within gender comparison. **(C)** Court: diachronic and contemporary comparison. **(D)** Letter: diachronic and contemporary comparison.

that the influence appears to go in the same direction, and with similar intensity, across both linguistic levels. Also, a grammatical influence of men over women is present, especially in the late period. This seems to indicate what has been shown in previous work: stylistic written features are adopted from men by women (cf. McIntosh, 1998, 205).

4.2.2.1. Micro-Analysis: Grammar

Inspecting features contributing to the period of change in the 1760s/70s for the contemporary model at the grammatical level (see **Figure 8D**), we see from **Figure 10** that women distinctively use verbal style (yellow lines), while men rely on nominal style (blue lines). Matching the results at the lexical level, women make use of a personal involved style marked by grammatical patterns of first person pronouns combined with mental verbs [pronoun.verb.pronoun (pnp.v.pnp), such as *I wish/think/hope you*] as well as modality and negation [pronoun.modalverb.not (pnp.mv.not) such as *I can not bear, I should not have, I could not write*]; see example (3). The few nominal patterns also reflect the involved style of writing with evaluative patterns that often include intensifiers [adverb.adjective.preposition (av.aj.prp) such as *very useful/rude/kind to, very strange/painful for*; noun.be.adverb

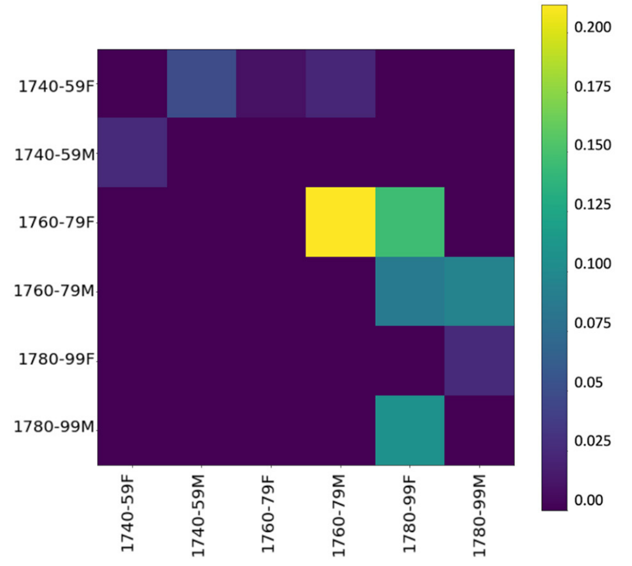


FIGURE 9 | Influencer heatmap of event cascades over time at the grammatical level (pos-trigrams).

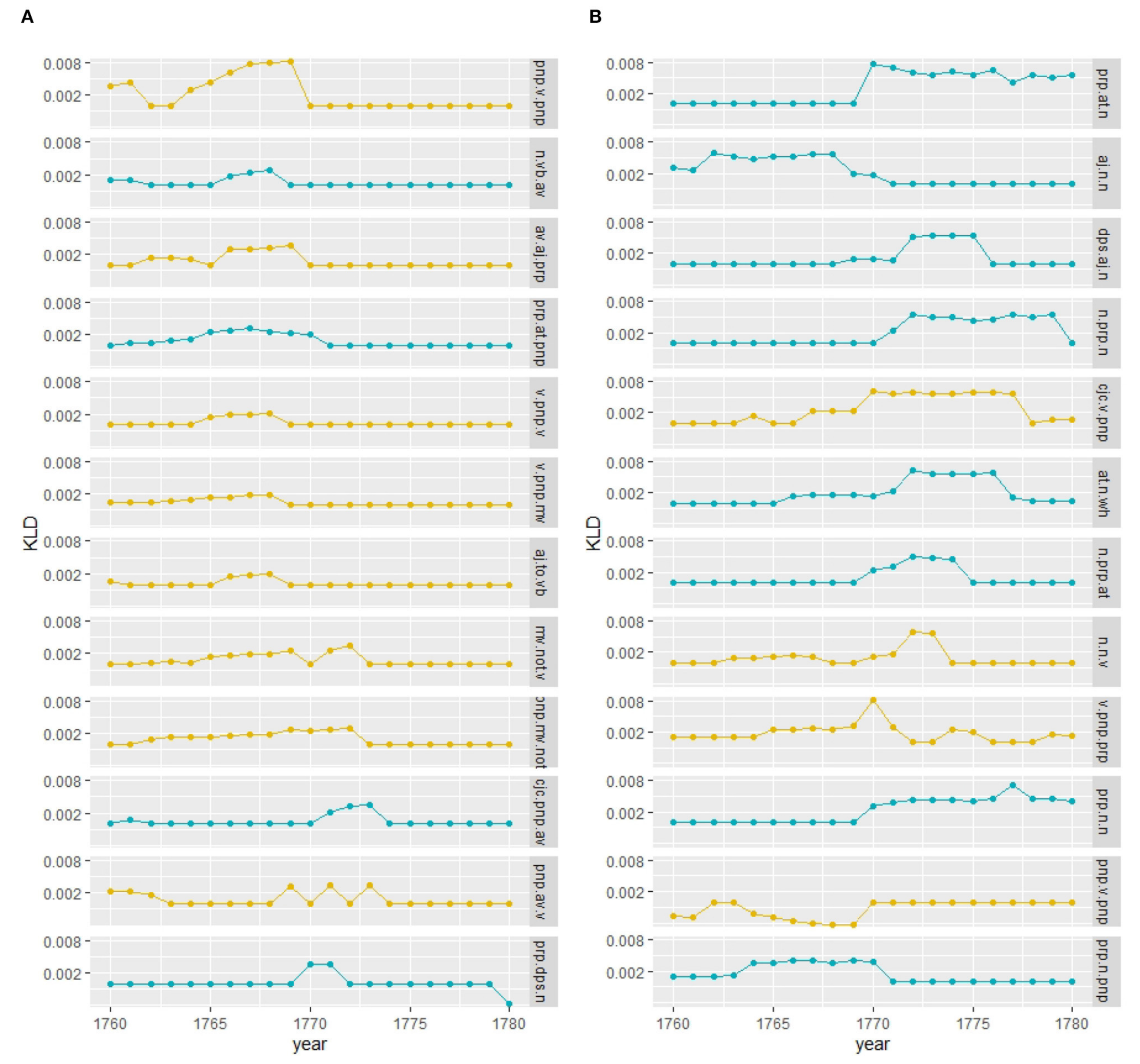


FIGURE 10 | Features contributing to period of change (from top to bottom showing highest to lowest contribution; blue: nominal trigrams, yellow: verbal trigrams)—contemporary model women vs. men (pos-trigrams). **(A)** Women. **(B)** Men. aj, adjective; av, adverb; cjc, conjunction; at, determiner; mv, modal verb; n, noun; dps, possessive *s*/pronoun; prp, preposition; pnp, pronoun; v, verb; vb, verb *be*; wh, wh-word.

(n.vb.av) such as *topic is ever (interesting), stomach is so (weak)*; preposition.possessive.noun (prp.dps.n) such as *in my opinion/mind*]; (4).

(3) for I have such a fixed depression upon my spirits, that **I cannot** raise them to any decent degree of Chearfulness, - when I have told you the Cause, **I think *you***, at least, will not wonder at the Effect.

(Frances Burney to her sister, Susanna, post - December 10, 1778; BURNEYF_011)

(4) Whoever is well acquainted with Venice must own it is the center of Pleasure, not so noisy, and **in my opinion** more refin'd than Paris. [...] He is singular both in his manner and Sentiments, yet I am apt to beleive if he meets with a sensible Wife, she may be **very happy with** him.

(Lady Mary Wortley Montagu to her daughter, Mary, Lady Bute, c. February 24, 1760; MONTAGU_192)

Men, instead, are distinguished by a very nominal and rather conventionalized style of writing using prepositional and

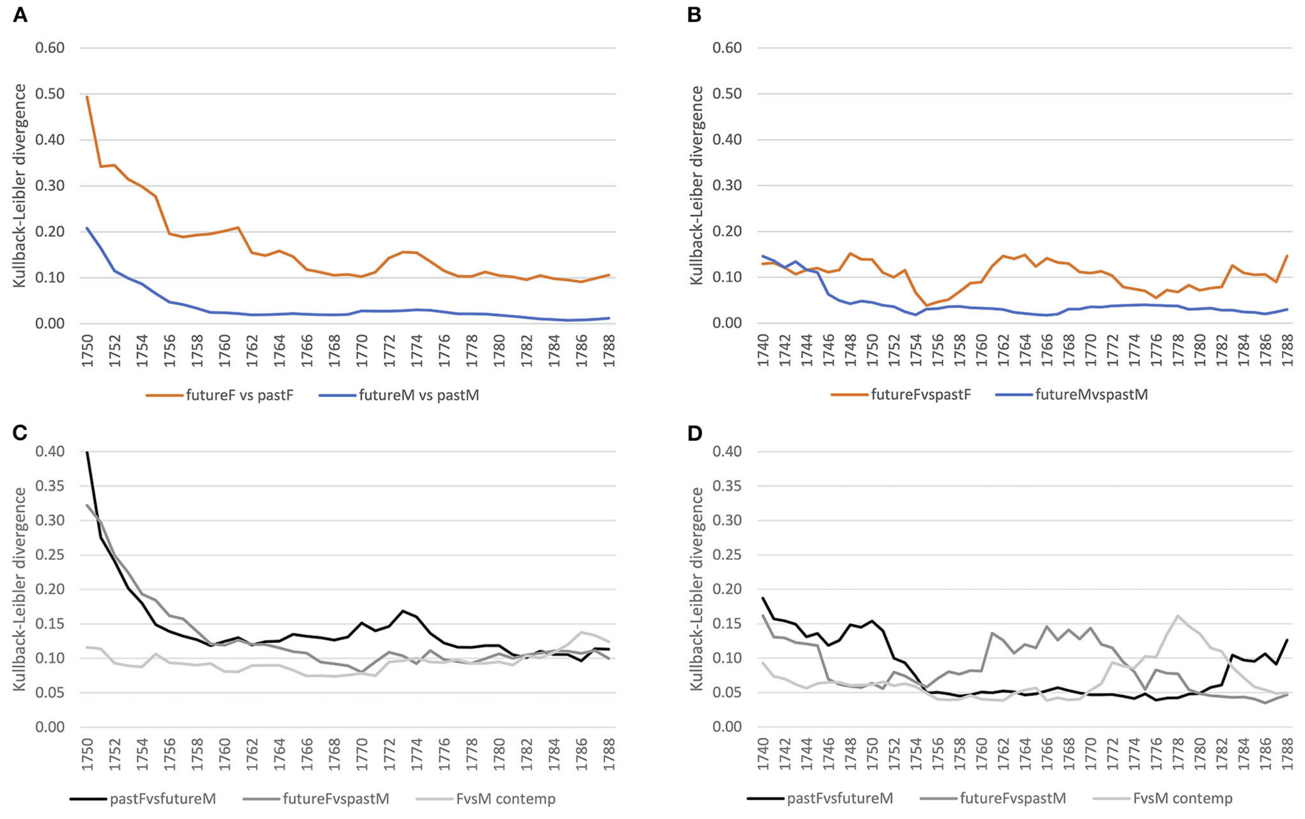


FIGURE 11 | Gender distinguished, diachronic, and contemporary models (court vs. letter) at the morphological level (suffixes). **(A)** Court: within gender comparison. **(B)** Letter: within gender comparison. **(C)** Court: diachronic and contemporary comparison. **(D)** Letter: diachronic and contemporary comparison.

compound patterns [e.g., preposition.determiner.noun (prp.at.n) as *to the queen, in the morning*, adjective.noun.noun (aj.n.n) as *small market town, tolerable drinking order*] used for narration of events, objects and places as well as verbal patterns of reporting [e.g., conjunction.verb.pronoun (cjc.v.pnp), such as *and said he, and told us*] and narration [e.g., verb.pronoun.preposition (v.pnp.prp) as *put him into, wrote it on, took us to* with material verbs; adverb.pronoun.verb (av.pnp.v) as *then they went, yesterday I sent, there we found* indicating place and time]. These patterns are illustrated in examples (5) and (6), written by two sons to their fathers; see also (2), where a husband narrates his activities using the adverb.pronoun.verb (av.pnp.v) pattern (*thence we saild*). Men (and boys; Pierce Taylor was a teenager at Eton) were often away from home and wrote about what they had seen and done, as well as general news of events in the places they were visiting. Even when women were traveling, like Lady Mary Wortley Montagu in (4), they tended to make their letters more about personal opinions and evaluation, or at least they used more explicit evaluative markers—compare (2), where Sir Roger Newdigate states his opinion about Lord Edgcombe's house but expresses it as a simple fact without hedging. These findings, then, match those at the lexical level, i.e., both linguistic levels reflect involved verbal language use of women strongly marked by modality, negation and evaluation vs. a more conventionalized nominal style of men in the informal setting focusing on narration of events, places and time.

(5) They **carried me to** the best houses **in the place**, shewed me whatever was worth seeing, and made several parties for me **in the country**.

(Edward Gibbon to his father, Edward Sr, May 31, 1763; GIBBON_013)

(6) **When we came into the Play Fields** the Sixth Form went to the Doctor **and said we** would all return if he would make us a Promise of Oblivion, He said No, Mr Roberts took Grenville and lock'd him up, on which we gathered round his House.

(Pierce Joseph Taylor to his father, Thomas, November 6, 1768; PIERCE_033)

4.2.3. Morphology

For the morphological level, we see similar converging trends within the same gender as for the lexical and grammatical levels for the court corpus (see **Figure 11A**). In the letter corpus, on the other hand, male language use converges, while female

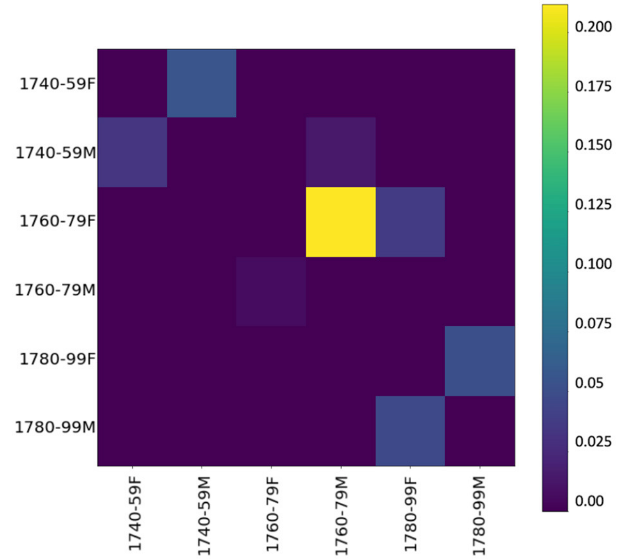


FIGURE 12 | Influencer heatmap of event cascades over time at the morphological level (suffixes).

language use fluctuates compared to the past, indicating periods of ongoing change (see **Figure 11B**).

Considering the contemporary models, female and male language use is relatively stable in the court corpus (see **Figure 11C**), while in the letter corpus, there is a peak around the 1780s. This period of change is also depicted in the diachronic models: future women diverge around the 1760s from past men and past women diverge from future men by the mid-1780s.

Our cascade model of influence (see **Figure 12**) confirms again that the influence of women over men is stronger than the other way around, especially in the period between the 1760s and the 1780s. Thus, starting from 1760 women introduce, rather than adopt, morphological innovations. Comparison with KLD results in **Figure 11D** shows that future female language use increasingly diverges from male language use of the past (dark gray line) in the 1750s. Thus, women seem to initiate a change adopted by men as shown in the cascade model.

In the last period, 1780–1799, divergence decreases in the contemporary model (see again **Figure 11D**). Here, unlike the other linguistic levels, the cascade model shows an influence of women and men on each other which also pertains to the last period.

4.2.3.1. Micro-Analysis: Morphology

A micro-analytic inspection reveals that the peak in KLD around the 1780s is largely due to the nominal suffix *-ness* (see **Figure 13**). While no gender difference has been found in its productivity in eighteenth-century correspondence as a whole (Säily, 2018a), in the final 20-year period of the corpus we do find women using it highly productively in family letters. The most productive users were published authors: Frances Burney, Mary Wollstonecraft, and Hester Lynch Piozzi. As shown by Säily (2018a, 214), *- ness* in the eighteenth century was increasingly being used in the sense "embodied attribute or trait," as in *your kindness*, and this change was led by women, potentially as part of the "feminization" of eighteenth-century culture (McIntosh, 2008, p. 231). Interestingly, while both men and women use the suffix in both positive and negative contexts, the types produced by women toward the end of the century are more skewed toward negative affect, as in (7).

(7) I hope God Almighty will preserve her to make us great Amends by her future Wisdom and Virtue for the Pain She now gives both to you and me by her **Grossness**, and her Contemptible Preference of the *Bon Ton and genteel Life* as She calls it, to every thing in this World and the next [...]

(Hester Lynch Piozzi to her daughter, Hester Maria Thrale, November 7, 1796; PIOZZI_061)

Another contributor to the peak around the 1780s is the nominal suffix *-ure*, again used most extensively by the same three women. As most of the types are borrowings from French, Latin or Italian with no base in English, this is not a case of increased productivity; however, the use of words with the foreign suffix may have been a way of signaling learnedness or sophistication, which would have been important to women like the intellectual Wollstonecraft or the Bluestockings Piozzi and Burney (Myers, 1990; Pohl and Schellenberg, 2003; Tieken-Boon van Ostade, 2010). Moreover, Piozzi married an Italian music teacher in the 1780s, while Burney later married a Frenchman, which may have influenced their use of *-ure* words, as in (8).

(8) M. d'Arblay has had a charming Letter from Comte Lally upon the ***brochure***—I intend also to enclose that, & dear Mr. Twining's, for your perusal, by Susanna.

(Frances Burney to her father, Charles Burney, c. January 27, 1794; BURNEYF_027)

4.2.4. Summarizing Temporal Dynamics of Gender-Specific Registerial Adaptation and Innovation

Overall, there is a converging trend in the formal setting of court proceedings across linguistic levels, pointing to registerial adaptation to formal conventions over time. Considering gender, the converging trend is steeper for women. In the informal setting of letters to family members, instead, across linguistic levels around the 1760s/70s/80s, women diverge from past language use of both women and men. Moreover, women lead changes which propagate to language use of later periods by men in particular. This clearly points to registerial innovation in the informal letter writing setting.

5. SUMMARY AND DISCUSSION

In this work, we take a long-term diachronic perspective, investigating gender-specific differences in language use. Our

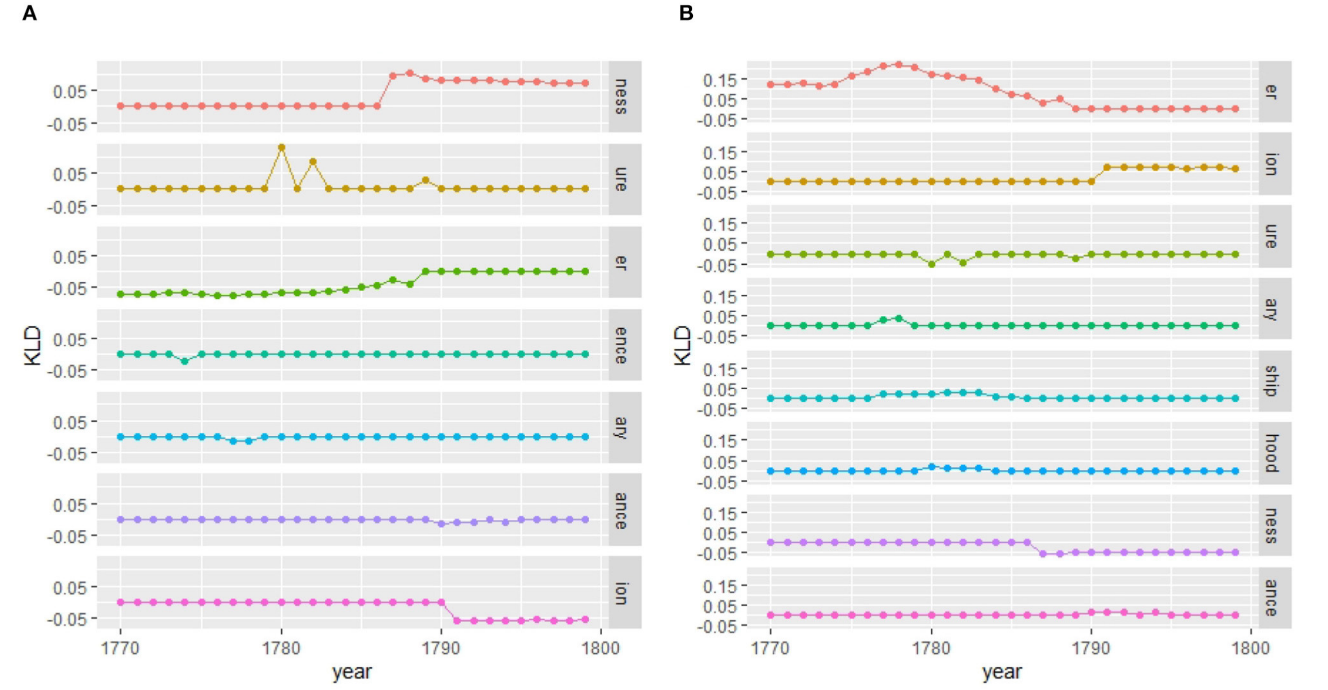


FIGURE 13 | Features contributing to period of change (from top to bottom showing highest to lowest contribution)—contemporary model women vs. men (suffixes). **(A)** Women. **(B)** Men.

focus is on women of the middle and upper classes as a social group in transition in this period, investigating change in language use across two different registers—one formal (court proceedings) and one informal (letters to family members). Thus, we consider two sociolinguistic factors (gender and register) involved in shaping the temporal dynamics of Late Modern British English in the 18th century. Computational methods have been used to model language use over time across three linguistic levels: lexis, grammar, and morphology.

We investigated two hypotheses: (a) *registerial adaptation* by middle- and upper-class women to formal conventions in the court setting, and, on the other hand, (b) *registerial innovation* in the informal setting of letters to family members. Findings have shown that underrepresented groups, such as women in the 18th century, can and do adapt to functional variation, such as registers, possibly triggered by external societal pressure. However, when no such pressure is at play, they create and shape new ways of using language and even lead changes which are then adopted not only by other women but also by men, who at that time enjoy a better social status.

These results could perhaps be seen to align with Labov's gender paradox, which states that "Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not" (Labov, 2001, p. 293). However, we cannot really state that women would conform *more* closely than men in the court setting, since we have established that they converge to men's language use, indicating that they are followers rather than leaders in the process of conventionalization there. The first part of the paradox would have been difficult to realize in the eighteenth-century courtroom since women had, by virtue of their gender, less access to the norms of this setting (see section 2 above). Therefore, we would not say that our results weaken Labov's claim *per se* but that its realization may depend on specific sociohistorical circumstances. Our results more unequivocally support the second part of the paradox as women innovate in the less tightly regulated setting of correspondence, which was also discovered by Nevalainen (2018, p. 258–259). In the eighteenth century, "letter-writing conventions became less formal, with their subject-matter including private as well as public matters, and letters were becoming an artistic, moral and intellectual literary form" (Somervell, 2011). Around the 1750s, the Bluestocking Society arose as a women's informal educational and social movement. Lady Mary Wortley Montagu's *Letters* from Turkey, for example, "were influential both as models of epistolary style and as anthropological works" (Somervell, 2011). Our results corroborate these findings in linguistic terms: from the 1760s onward, women initiate changes across linguistic levels diverging from past language use of both women and men—changes which are subsequently adopted by men.

While we have focused more on the informal setting in this paper, it would be interesting to also more deeply investigate how gender-specific change in language use has propagated within more formal registers, such as court proceedings (cf. Degaetano-Ortlieb, 2018) or scientific writing (cf. Degaetano-Ortlieb and Teich, 2019), given enough gender-annotated data. Furthermore, considering that conventionalization seems essential in language change as a precondition for innovation (Bybee, 2010; Schmid, 2015; De Smet, 2016; Teich et al., 2021), it would be worth studying the interplay between convention and innovation from a gender perspective. In addition, given enough data, a network analysis would be intriguing to trace the propagation of innovation across individuals (see, e.g., Sairio, 2009).

Our study offers two main methodological contributions to the analysis of long-term diachronic data, adapting computational modeling to form novel ways of inspecting long-term temporal dynamics of language use. For detection of change, we use diachronic periodization based on Kullback-Leibler divergence, allowing us to determine when changes occur (rather than using pre-defined periods for comparison), using a wide range of features (avoiding pre-selection bias) across linguistic levels, and to derive relevant features of variation in language use from the data at hand. The application of event cascades based on the Multivariate Hawkes Process (usually employed in sociolinguistics to model turn-taking interactions) on long-term diachronic data allows us to model influencer groups of women and men over time. Our results conform with a long-term assumption of women being involved in leading registerial change over time in more informal settings.

A major challenge that has to be faced in diachronic analysis is the representativeness of corpus data. While diachronic corpora are extremely valuable resources and the compilers of such corpora have undertaken immense effort to create these resources in the best way possible, the fact remains that representativeness cannot be achieved as fully as for contemporary synchronic studies (cf. Gries and Hilpert, 2008). Diachronic data is ultimately a finite sample constrained by past data availability. Nevertheless, our findings from the perspective of sociolinguistics corroborate findings across other disciplines, such as literary studies and history, showing how computational sociolinguistic work adds to the scientific endeavor of better understanding the temporal dynamics of change. The methodology of using Kullback-Leibler Divergence to detect change has already been applied to other diachronic corpora and has shown its validity on smaller ones as well.[13] Regardless of size, spelling variation is also an issue for diachronic corpora. Better methods of spelling normalization are currently being developed (e.g., Hämäläinen et al., 2019), an endeavor that should be pursued further. Future methodological development in dealing with these small but complex datasets should perhaps especially focus on issues of potential bias and outliers, so that they could be alleviated, and so that human analysts would be alerted to particularly sparse or skewed data in specific time periods. In diachronic research, computational analysis should always be complemented with qualitative human analysis, contextualization, and interpretation, as we have striven to do here.

[13] For example, KLD has been applied to the *Royal Society Corpus* (RSC; Kermes et al., 2016; Fischer et al., 2020) with ∼32 million tokens, the *Corpus of Late Modern English Texts* (CLMET; Diller et al., 2010) with ∼40 million tokens (see Degaetano-Ortlieb and Teich, 2019) as well as to a small corpus of history texts (CHET; Moskowich et al., 2019) with 40 texts amounting to ∼500,000 tokens (see Degaetano-Ortlieb et al., 2019c).

DATA AVAILABILITY STATEMENT

The *Old Bailey Corpus* v.2.0 is available at: https://hdl.handle.net/11858/00-246C-0000-0023-8CFB-2 (persistent handle). Owing to copyright reasons, access to the *Tagged Corpus of Early English Correspondence Extension* as a whole is restricted to on-site use at the University of Helsinki; however, parts can be made available for research purposes upon request. Data and scripts are made available at: https://github.com/degaetano-ortlieb/frontiersCompSocioLingRegisterGender.

AUTHOR CONTRIBUTIONS

SD-O and TS contributed the conception and design of the study. SD-O wrote the first draft of the manuscript, provided the background on register studies, introduced the data-driven periodization by KLD, conducted the KLD analysis, and interrelated the KLD, cascade, and micro-analytical results. TS provided the historical sociolinguistic background and interpretation, information on the TCEECE, data cleaning for the suffixes with SD-O, and micro-analytic inspection of the study, partly with SD-O at the lexical and grammatical levels. YB ran the preliminary Hawkes Process investigations on the TCEECE dataset, adapted the event cascade framework to the long-term gender influence scenario, and conducted the event cascade analyses. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft, Funder Id: http://dx.doi.org/10.13039/501100001659, Grant Number: SFB1102: Information Density and Linguistic Encoding, and by the Academy of Finland, Grants 293009 and 323390.

ACKNOWLEDGMENTS

We would like to thank Lassi Saario and Juho Ruohonen for assistance in cleaning up the suffix data, and Yasmin Arbelaiz for assistance in bibliography notation. We were also indebted to Terttu Nevalainen, Andrew Piper, and Elke Teich for reading a previous version of this manuscript and to Badr Abdullah and Marius Mosbach for valuable feedback on the more technical part. And last but not least thanks to the reviewers for their valuable comments, which improved the manuscript further.

REFERENCES

Allan, B. (1976). Ordinal-scaled variables and multivariate analysis: comment on Hawkes. *Am. J. Sociol.* 81, 1498–1500. doi: 10.1086/226239

Argamon, S. (2019). Register in computational language research. *Regist. Stud.* 1, 100–135. doi: 10.1075/rs.18015.arg

Argamon, S., Dodick, J., and Chase, P. (2008). Language use reflects scientific methodology: a corpus-based study of peer-reviewed journal articles. *Scientometrics* 75, 203–238. doi: 10.1007/s11192-007-1768-y

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text* 23, 321–346. doi: 10.1515/text.2003.014

Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal. *Appl. Linguist*. 13, 337–374. doi: 10.1093/applin/13.4.337

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *J. Sociolinguist*. 18, 135–160. doi: 10.1111/josl.12080

Baron, A. (2011a). *VARD 2*. Computer program. Lancaster: Lancaster University. Available online at: http://ucrel.lancs.ac.uk/vard/

Baron, A. (2011b). *Dealing with spelling variation in early modern English texts* (Ph.D. thesis), Lancaster University, Lancaster, United Kingdom.

Barron, A. T. J., Huang, J., Spang, R. L., and DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proc. Natl. Acad. Sci. U.S.A*. 115, 4607–4612. doi: 10.1073/pnas.1717729115

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A., and Romanov, M. (2019). Studying the history of the Arabic language: language technology and a large-scale historical corpus. *Lang. Resour. Eval*. 53, 771–805. doi: 10.1007/s10579-019-09460-w

Bermudez-Otero, R., and Trousdale, G. (2012). "Cycles and continua: on unidirectionality and gradualness in language change," in *The Oxford Handbook of the History of English*, eds T. Nevalainen and E. C. Traugott (Oxford: Oxford University Press), 691–720. doi: 10.1093/oxfordhb/9780199922765.013.0059

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511621024

Biber, D., and Burges, J. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. *J. English Linguist*. 28, 21–37. doi: 10.1177/00754240022004857

Biber, D., and Conrad, S. (2001). "Register variation: a corpus approach," in *The Handbook of Discourse Analysis*, eds D. Schiffrin, D. Tannen, and H. E. Hamilton (Chichester: Wiley Online Library), 175–196. doi: 10.1002/9780470753460.ch10

Biber, D., and Finegan, E. (1997). "Diachronic relations among speech-based and written registers in English," in *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, eds T. Nevalainen and L. Kahlas-Tarkka (Helsinki: Société Néophilologique), 253–276.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., and Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: a data-driven approach. *Front. Artif. Intell*. 3:73. doi: 10.3389/frai.2020.00073

Bochkarev, V., Solovyev, V. D., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* 11, 1–8. doi: 10.1098/rsif.2014.0841

Bod, R., Hay, J., and Jannedy, S. (Eds.). (2003). *Probabilistic Linguistics*. Cambridge; London: MIT Press. doi: 10.7551/mitpress/5582.001.0001

Broccias, C. (2012). "The syntax-lexicon continuum," in *The Oxford Handbook of the History of English*, eds T. Nevalainen and E. C. Traugott (Oxford: Oxford University Press), 735–747. doi: 10.1093/oxfordhb/9780199922765.013.0061

Bybee, J. L. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511750526

Carr, R. (2009). *Review of Women and Enlightenment in Eighteenth-Century Britain (Review No. 831)*. Reviews in History. Available online at: https://reviews.history.ac.uk/review/831

Claridge, C. (2012). "Registers, genres and the standard: some thoughts on the corpus-linguistic documentation of the 18th century," in *Codification, Canons and Curricula: Description and Prescription in Language and Literature*, eds A. Schröder, U. Busse, and R. Schneider (Bielefeld: Aisthesis), 79–92.

CLAWS (1994). *Computer Program*. Developed by UCREL at Lancaster University. Lancaster: Lancaster University.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Abingdon: Routledge. doi: 10.4324/9780203830901

Culpeper, J., and Kytö, M. (2010). *Early Modern English Dialogues: Spoken Interaction as Writing*. Studies in English Language. Cambridge: Cambridge University Press.

Daw, A., Castellanos, A., Yom-Tov, G. B., Pender, J., and Gruendlinger, L. (2020). The co-production of service: modeling service times in contact centers using Hawkes processes. *arXiv* 2004.07861. doi: 10.2139/ssrn.3817130

De Smet, H. (2006). A corpus of Late Modern English texts. *ICAME J*. 29, 69–82. Available online at: http://icame.uib.no/ij29/

De Smet, H. (2016). How gradual change progresses: the interaction between convention and innovation. *Lang. Variat. Change* 28, 83–102. doi: 10.1017/S0954394515000186

Degaetano-Ortlieb, S. (2015). *Evaluative meaning in scientific writing: macro- and micro-analytic perspectives using data mining* (Ph.D. thesis), Universität des Saarlandes, Saarbrücken, Germany.

Degaetano-Ortlieb, S. (2018). "Stylistic variation over 200 years of court proceedings according to gender and social class," in *Proceedings of the 2nd Workshop on Stylistic Variation Collocated With NAACL HLT 2018* (New Orleans, LA: Association for Computational Linguistics), 1–10. doi: 10.18653/v1/W18-1601

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., and Teich, E. (2019a). "An information-theoretic approach to modeling diachronic change in scientific English," in *From Data to Evidence in English Language Research, Language and Computers*, eds C. Suhr, T. Nevalainen, and I. Taavitsainen (Leiden: Brill), 258–281. doi: 10.1163/9789004390652_012

Degaetano-Ortlieb, S., Krielke, P., Scheurer, F., and Teich, E. (2019b). "A diachronic perspective on efficiency in language use: *that*-complement clause in academic writing across 300 years," in *Proceedings of the 10th International Corpus Linguistics Conference* (Cardiff).

Degaetano-Ortlieb, S., Menzel, K., and Teich, E. (2019c). "Typical linguistic patterns of English history texts from the eighteenth to the nineteenth century," in *Writing History in Late Modern English: Explorations of the Coruña Corpus*, eds I. Moskowich, B. Crespo, L. Puente-Castelo, and L. M. Monaco (Amsterdam: John Benjamins), 58–81. doi: 10.1075/z.225.04deg

Degaetano-Ortlieb, S., and Piper, A. (2019). "The scientization of literary study," in Proceedings of the 3nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at NAACL, Number W19-25 in ACL Anthology (Minneapolis, MN: Association for Computational Linguistics), 18–28. doi: 10.18653/v1/W19-2503

Degaetano-Ortlieb, S., and Strötgen, J. (2018). "Diachronic variation of temporal expressions in scientific writing through the lens of relative entropy," in *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Proceedings, Volume 10713 of Lecture Notes in Computer Science*, eds G. Rehm and T. Declerck (Berlin: Springer International Publishing), 259–275. doi: 10.1007/978-3-319-73706-5_22

Degaetano-Ortlieb, S., and Teich, E. (2016). "Information-based modeling of diachronic linguistic change: from typicality to productivity," in *Proceedings of the 10th SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, and Humanities at ACL* (Berlin: Association for Computational Linguistics), 165–173. doi: 10.18653/v1/W16-2121

Degaetano-Ortlieb, S., and Teich, E. (2017). "Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns," in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at ACL* (Vancouver, BC: ACL), 68–77. doi: 10.18653/v1/W17-2209

Degaetano-Ortlieb, S., and Teich, E. (2018). "Using relative entropy for detection and analysis of periods of diachronic linguistic change," in *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING* (Santa Fe, NM: Association for Computational Linguistics), 22–33.

Degaetano-Ortlieb, S., and Teich, E. (2019). Toward an optimal code for communication: the case of scientific English. *Corpus Linguist. Linguist. Theory* 1–33. doi: 10.1515/cllt-2018-0088

Diller, H. J., De Smet, H., and Tyrkkö, J. (2010). A European database of descriptors of English electronic texts. *Eur. English Messenger* 19, 29–35. Available online at: https://essenglish.org/messenger/back/issue-xix2-autumn-2010/

Dutta, H. S., Dutta, V. R., Adhikary, A., and Chakraborty, T. (2020). HawkesEye: detecting fake retweeters using Hawkes process and topic modeling. *IEEE Trans. Inform. Forensics Security* 15, 2667–2678. doi: 10.1109/TIFS.2020.2970601

Eisenstein, J. (2015). "Identifying regional dialects in on–line social media," in *Handbook of Dialectology*, eds C. Boberg, J. Nerbonne, and D. Watt (Chichester: Wiley), 368–383. doi: 10.1002/9781118827628.ch21

Eisenstein, J. (2019). "Measuring and modeling language change," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (Minneapolis, MN: Association for Computational Linguistics), 9–14.

Eisenstein, J., Smith, N. A., and Xing, E. P. (2011). "Discovering sociolinguistic associations with structured sparsity," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, OR: Association for Computational Linguistics), 1365–1374.

Emsley, C., Hitchcock, T., and Shoemaker, R. (2018a). "Historical background—gender in the proceedings," in *Old Bailey Proceedings Online, Version 7.0*. Available online at: https://www.oldbaileyonline.org

Emsley, C., Hitchcock, T., and Shoemaker, R. (2018b). "The proceedings—the value of the proceedings as a historical source," in *Old Bailey Proceedings Online, Version 7.0*. Available online at: https://www.oldbaileyonline.org

Evert, S. (2005). *The CQP Query Language Tutorial*. CWB version 2.2.b90. Stuttgart: IMS Stuttgart.

Fankhauser, P., Knappen, J., and Teich, E. (2014). "Exploring and visualizing variation in language resources," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)* (Reykjavik: European Language Resources Association), 4125–4128.

Ferguson, C. A. (1994). "Dialect, register, and genre: working assumptions about conventionalization," in *Sociolinguistic Perspectives on Register*, eds D. Biber and E. Finegan (Oxford: Oxford University Press), 15–30.

Ferrara, K., Brunner, H., and Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Commun.* 8, 8–34. doi: 10.1177/0741088391008001002

Ficler, J., and Goldberg, Y. (2017). "Controlling linguistic style aspects in neural language generation," in *Proceedings of the Workshop on Stylistic Variation* (Copenhagen: Association for Computational Linguistics), 94–104. doi: 10.18653/v1/W17-4912

Fischer, S., Knappen, J., Menzel, K., and Teich, E. (2020). "The Royal Society Corpus 6.0: providing 300+ years of scientific writing for humanistic study," in *Proceedings of the Language Resources and Evaluation Conference (LREC)* (Marseille: European Language Resources Association), 794–802.

Freund, L., Clarke, C. L. A., and Toms, E. G. (2006). "Towards genre classification for IR in the workplace," in *Proceedings of the 1st International Conference on Information Interaction in Context, IIiX* (New York, NY: Association for Computing Machinery), 30–36. doi: 10.1145/1164820.1164829

Giesbrecht, E., and Evert, S. (2009). "Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus," in *Proceedings of the Fifth Web as Corpus Workshop* (Saarbruecken), 27–35.

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., et al. (2016). "The social dynamics of language change in online networks," in *International Conference on Social Informatics* (Cham: Springer), 41–57. doi: 10.1007/978-3-319-47880-7_3

Gries, S., and Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3, 59–81. doi: 10.3366/E1749503208000075

Gries, S., and Hilpert, M. (2010). Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Lang. Linguist.* 14, 59–81. doi: 10.1017/S1360674310000092

Halliday, M. (1988). "On the language of physical science," in *Registers of Written English: Situational Factors and Linguistic Features*, ed M. Ghadessy (London: Pinter), 162–177.

Halliday, M. (1989). *Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M. (2004). *An Introduction to Functional Grammar*. London: Hodder Education.

Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., and Mäkelä, E. (2019). "Revisiting NMT for normalization of early English letters," in *Proceedings of the Third Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL-2019)*, eds B. Alex, S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz (Stroudsburg, PA: Association for Computational Linguistics), 71–75. doi: 10.18653/v1/W19-2509

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *J. R. Statist. Soc. B Methodol.* 33, 438–443. doi: 10.1111/j.2517-6161.1971.tb01530.x

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quant. Finance* 18, 193–198. doi: 10.1080/14697688.2017.1403131

Hay, D., and Rogers, N. (1997). *Eighteenth-Century English Society: Shuttles and Swords*. Oxford: Oxford University Press.

Herring, S. C., and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *J. Sociolinguist.* 10, 439–459. doi: 10.1111/j.1467-9841.2006.00287.x

Heylighen, F., and Dewaele, J. M. (2002). Variation in the contextuality of language: an empirical measure. *Foundat. Sci.* 7, 293–340. doi: 10.1023/A:1019661126744

Huber, M. (2007). "The Old Bailey Proceedings 1674–1834: evaluating and annotating a corpus of 18th- and 19th-century spoken English," in *Annotating Variation and Change, Number 1 in Studies in Variation, Contacts and Change in English*, eds A. Meurman-Solin and A. Nurmi (Helsinki: VARIENG).

Huber, M., Nissel, M., and Puga, K. (2016). *Old Bailey Corpus 2.0*. Giessen: Justus Liebig University Giessen. hdl:11858/00–246C-0000–0023-8CFB-2.

Hughes, J. M., Foti, N. J., Krakauer, D. C., and Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7682–7686. doi: 10.1073/pnas.1115407109

Jhamtani, H., Gangal, V., Hovy, E., and Nyberg, E. (2017). "Shakespearizing modern language using copy-enriched sequence to sequence models," in *Proceedings of the Workshop on Stylistic Variation* (Copenhagen: Association for Computational Linguistics), 10–19. doi: 10.18653/v1/W17-4902

Ji, M. (2010). A corpus-based study of lexical periodization in historical Chinese. *Liter. Linguist. Comput.* 25, 199–213. doi: 10.1093/llc/fqq002

Kaislaniemi, S. (2018). "The Corpus of Early English Correspondence Extension (CEECE)," in *Patterns of Change in 18th-Century English: A Sociolinguistic Approach, Number 8 in Advances in Historical Sociolinguistics*, eds T. Nevalainen, M. Palander-Collin, and T. Säily (Amsterdam: John Benjamins), 45–59. doi: 10.1075/ahs.8.04kai

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., and Teich, E. (2016). "The Royal Society Corpus: from uncharted data to corpus," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)* (Portorož: European Language Resources Association), 1928–1931.

Klingenstein, S., Hitchcock, T., and DeDeo, S. (2014). The civilizing process in London's Old Bailey. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9419–9424. doi: 10.1073/pnas.1405984111

Kopaczyk, J. (Ed.). (2013). *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles. Oxford Studies in Language and Law*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199945153.001.0001

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22, 79–86. doi: 10.1214/aoms/1177729694

Kytö, M. (1993). Third-person present singular verb inflection in early British and American English. *Lang. Variation Change* 5, 113–139. doi: 10.1017/S0954394500001447

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Language in Society. Oxford: Blackwell Publishers.

Labov, W. (2001). *Principles of Linguistic Change, Volume 2: Social Factors*. Language in Society. Oxford: Blackwell Publishers.

Linderman, S., and Adams, R. (2014). "Discovering latent network structure in point process data," in *Proceedings of the 31st International Conference on Machine Learning* (Beijing: PLMR), 1413–1421.

Markus, M. (2001). The development of prose in Early Modern English in view of the gender question: using grammatical idiosyncracies of 15th and 17th century letters. *Eur. J. English Stud.* 5, 181–196. doi: 10.1076/ejes.5.2.181.7305

Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam: John Benjamins. doi: 10.1075/z.59

Mauch, M., MacCallum, R. M., Levy, M., and Leroi, A. M. (2015). The evolution of popular music: USA 1960–2010. *R. Soc. Open Sci.* 2:150081. doi: 10.1098/rsos.150081

McIntosh, C. (1998). *The Evolution of English Prose 1700–1800: Style, Politeness, and Print Culture*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511582790

McIntosh, C. (2008). "British English in the long eighteenth century (1660–1830)," in *A Companion to the History of the English Language*, eds H. Momma and M. Matto (Chichester: Wiley-Blackwell), 228–234. doi: 10.1002/9781444302851.ch22

Milroy, J., and Milroy, L. (1991). *Authority in Language*. London: Routledge.

Morato, J., Llorens, J., Genova, G., and Moreiro, J. A. (2003). Experiments in discourse analysis impact on information classification and retrieval algorithms. *Inform. Process. Manage*. 39, 825–851. doi: 10.1016/S0306-4573(02)00081-X

Moskowich, I., Crespo, B., Puente-Castelo, L., and Monaco, L. (2019). *Writing History in Late Modern English: Explorations of the Coruña Corpus*. Amsterdam: John Benjamins. doi: 10.1075/z.225

Myers, S. H. (1990). *The Bluestocking Circle: Women, Friendship, and the Life of the Mind in Eighteenth-Century England*. Oxford: Clarendon Press.

Nevalainen, T. (2018). "A wider sociolinguistic perspective," in *Patterns of Change in 18th-Century English: A Sociolinguistic Approach, Number 8 in Advances in Historical Sociolinguistics*, eds T. Nevalainen, M. Palander-Collin, and T. Säily (Amsterdam: John Benjamins), 255–270. doi: 10.1075/ahs.8.16nev

Nevalainen, T., Palander-Collin, M., and Säily, T. (Eds.). (2018). *Patterns of Change in 18th-Century English: A Sociolinguistic Approach. Number 8 in Advances in Historical Sociolinguistics*. Amsterdam: John Benjamins. doi: 10.1075/ahs.8.01nev

Nevalainen, T., and Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Longman Linguistics Library. London: Pearson Education.

Nevalainen, T., and Raumolin-Brunberg, H. (Eds.). (1996). *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence. Number 15 in Language and Computers: Studies in Practical Linguistics*. Amsterdam: Rodopi.

Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M., et al. (1998–2006). *CEEC, Corpora of Early English Correspondence*. Helsinki: Department of Modern Languages, University of Helsinki.

Nevalainen, T., and Traugott, E. C. (Eds.). (2012). *The Oxford Handbook of the History of English*. New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199922765.001.0001

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discour. Process*. 45, 211–236. doi: 10.1080/01638530802073712

Nguyen, D., Dogruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: a survey. *CoRR* abs/1508.07544. doi: 10.1162/COLI_a_00258

Nowson, S., Oberlander, J., and Gill, A. J. (2005). "Weblogs, genres and individual differences," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1666–1671. Available online at: https://escholarship.org/uc/item/8bg0t4c6

O'Brien, K. (2009). *Women and Enlightenment in Eighteenth-Century Britain*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511576317

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books Corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10:e0137041. doi: 10.1371/journal.pone.0137041

Petré, P., and Anthonissen, L. (2020). Individuality in complex systems: a constructionist approach. *Cogn. Linguist*. 31, 185–212. doi: 10.1515/cog-2019-0033

Petré, P., and Van de Velde, F. (2018). The real-time dynamics of the individual and the community in grammaticalization. *Language* 94, 867–901. doi: 10.1353/lan.2018.0056

Pohl, N., and Schellenberg, B. A. (Eds.). (2003). *Reconsidering the Bluestockings*. San Marino, CA: Huntington Library.

Popescu, O., and Strapparava, C. (2013). "Behind the times: detecting epoch changes using large corpora," in *International Joint Conference on Natural Language Processing* (Nagoya: Asian Federation of Natural Language Processing), 347–355.

Raumolin-Brunberg, H., and Nevalainen, T. (2007). "Historical sociolinguistics: the Corpus of Early English Correspondence," in *Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases*, eds J. C. Beal, K. P. Corrigan, and H. L. Moisl (Houndsmills: Palgrave Macmillan), 148–171. doi: 10.1057/9780230223202_7

Ravid, D., and Berman, R. (2009). Developing linguistic register across text types: the case of modern Hebrew. *Pragmat. Cogn*. 17, 108–145. doi: 10.1075/pc.17.1.04rav

Ravid, D., and Tolchinsky, L. (2002). Developing linguistic literacy: a comprehensive model. *J. Child Lang*. 29, 417–447. doi: 10.1017/S0305000902005111

Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *Int. J. Corpus Linguist*. 2, 133–152. doi: 10.1075/ijcl.2.1.07ray

Reiter, E., and Williams, S. (2010). "Generating texts in different styles," in *The Structure of Style: Algorithmic Approaches to Manner and Meaning*, eds S. Argamon, K. Burns, and S. Dubnov (Heidelberg: Springer), 59–78. doi: 10.1007/978-3-642-12337-5_4

Saario, L., and Säily, T. (2020). *POS Tagging the CEECE. A Manual to Accompany the Tagged Corpus of Early English Correspondence Extension (TCEECE)*. Helsinki: VARIENG. Available online at: https://varieng.helsinki.fi/CoRD/corpora/CEEC/tceece_doc.html

Säily, T. (2016). Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguist. Linguist. Theory* 12, 129–151. doi: 10.1515/cllt-2015-0064

Säily, T. (2018a). "Change or variation? Productivity of the suffixes *-ness* and *-ity*," in *Patterns of Change in 18th-century English: A Sociolinguistic Approach, Number 8 in Advances in Historical Sociolinguistics*, eds T. Nevalainen, M. Palander-Collin, and T. Säily (Amsterdam: John Benjamins), 197–218.

Säily, T. (2018b). "Conservative and progressive individuals," in *Patterns of Change in 18th-Century English: A Sociolinguistic Approach, Number 8 in Advances in Historical Sociolinguistics*, eds T. Nevalainen, M. Palander-Collin, and T. Säily (Amsterdam: John Benjamins), 235–242. doi: 10.1075/ahs.8.14sai

Säily, T., and Mäkelä, E. (2019). *The OED and Historical Text Collections: Discovering New Words*. Oxford English Dictionary Webinar Series. Oxford: Oxford University Press. Available online at: https://www.youtube.com/watch?v=tL5t_-IPacE

Säily, T., Nevalainen, T., and Siirtola, H. (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Liter. Linguist. Comput*. 26, 167–188. doi: 10.1093/llc/fqr004

Säily, T., Nurmi, A., Palander-Collin, M., and Auer, A. (Eds.). (2017a). *Exploring Future Paths for Historical Sociolinguistics. Number 7 in Advances in Historical Sociolinguistics*. Amsterdam: John Benjamins. doi: 10.1075/ahs.7

Säily, T., Vartiainen, T., and Siirtola, H. (2017b). "Exploring part-of-speech frequencies in a sociohistorical corpus of English," in *Exploring Future Paths for Historical Sociolinguistics, Number 7 in Advances in Historical Sociolinguistics*, eds T. Säily, A. Nurmi, M. Palander-Collin, and A. Auer (Amsterdam: John Benjamins), 23–52. doi: 10.1075/ahs.7.02sai

Sairio, A. (2009). "Methodological and practical aspects of historical network analysis: a case study of the Bluestocking letters," in *The Language of Daily Life in England (1400–1800)*, eds A. Nurmi, M. Nevala, and M. Palander-Collin (Amsterdam: John Benjamins), 107–135. doi: 10.1075/pbns.183.08sai

Schmid, H. -J. (2015). A blueprint of the entrenchment-and-conventionalization model. *Yearb. German Cogn. Linguist. Assoc*. 3, 3–26. doi: 10.1515/gcla-2015-0002

Sharoff, S., Wu, Z., and Markert, K. (2010). "The web library of Babel: evaluating genre collections," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (Valletta: European Language Resources Association), 3063–3070.

Somervell, T. (2011). *Public and Private, Real and Fictional: The Rise of Women's Letter-Writing in the Eighteenth Century*. Bluestocking: Online Journal for Women's History. Available online at: https://blue-stocking.org.uk/2011/03/01/public-and-private-real-and-fictional-the-rise-of-womens-letter-writing-in-the-eighteenth-century/

Tagliamonte, S. A. (2009). "*Be + like*: the new quotative in English," in *The New Sociolinguistics Reader*, eds N. Coupland and A. Jaworski (Basingstoke: Palgrave Macmillan), 75–91. doi: 10.1007/978-1-349-92299-4_6

Tagliamonte, S. A. (2012). *Variationist Sociolinguistics: Change, Observation, Interpretation. Number 40 in Language in Society*. Malden, MA: Wiley-Blackwell.

TCEECE (2018). *Tagged Corpus of Early English Correspondence Extension*. Helsinki: Department of Modern Languages, University of Helsinki. Available online at: https://varieng.helsinki.fi/CoRD/corpora/CEEC/

Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H., and Lapshinova-Koltunski, E. (2016). The linguistic construal of disciplinarity: a data mining approach using register features. *J. Assoc. Inform. Sci. Technol*. 67, 1668–1678. doi: 10.1002/asi.23457

Teich, E., Degaetano-Ortlieb, S., Kermes, H., and Lapshinova-Koltunski, E. (2013). "Scientific registers and disciplinary diversification: a comparable corpus approach," in *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora* (Sofia: Association for Computational Linguistics), 59–68.

Teich, E., and Fankhauser, P. (2010). "Exploring a corpus of scientific texts using data mining," in *Corpus-Linguistic Applications: Current Studies, New Directions*, eds S. Gries, S. Wulff, and M. Davies (Amsterdam; New York, NY: Rodopi), 233–247. doi: 10.1163/9789042028012_016

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., and Bizzoni, Y. (2021). Less is more/more diverse: on the communicative utility of linguistic conventionalization. *Front. Commun*. 5:142. doi: 10.3389/fcomm.2020.620275

Tieken-Boon van Ostade, I. (2010). "Eighteenth-century women and their norms of correctness," in *Eighteenth-Century English: Ideology and Change, Studies in English Language*, ed R. Hickey (Cambridge: Cambridge University Press), 59–72. doi: 10.1017/CBO9780511781643.005

Ure, J. (1971). Lexical density and register differentiation. *Contemp. Educ. Psychol*. 5, 96–104.

Ure, J. (1982). Introduction: approaches to the study of register range. *Int. J. Sociol. Lang*. 35, 5–23. doi: 10.1515/ijsl.1982.35.5

van Hulle, D., and Kestemont, M. (2016). Periodizing Samuel Beckett's works: a stylochronometric approach. *Style* 50, 172–202. doi: 10.1353/sty.2016.0003

Vartiainen, T., Säily, T., and Hakala, M. (2013). "Variation in pronoun frequencies in early English letters: gender-based or relationship-based?" in *Ex Philologia Lux: Essays in Honour of Leena Kahlas-Tarkka, Number XC in Mémoires de la Société Néophilologique de Helsinki*, eds J. Tyrkkö, O. Timofeeva, and M. Salenius (Helsinki: Société Néophilologique), 233–255.

Ylivuori, S. (2019). *Women and Politeness in Eighteenth-Century England. Routledge Studies in Eighteenth-Century Cultures and Societies*. New York, NY: Routledge. doi: 10.4324/9780429454431

Yuan, B., Li, H., Bertozzi, A. L., Brantingham, P. J., and Porter, M. A. (2019). Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM J. Math. Data Sci*. 1, 356–382. doi: 10.1137/18M1226993

Zhai, C., and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inform. Syst*. 22, 179–214. doi: 10.1145/984321.984322

Zhang, W., Bu, F., Owens-Oas, D., Heller, K., and Zhu, X. (2018). Learning root source with marked multivariate Hawkes processes. *arXiv* 1809. 03648.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 04 June 2021
doi: 10.3389/frai.2021.668035

Linking Linguistic and Geographic Distance in Four Semantic Domains: Computational Geo-Analyses of Internal and External Factors in a Dialect Continuum

John L. A. Huisman[1]*, *Karlien Franco*[2] *and Roeland van Hout*[1]

[1]Centre for Language Studies, Radboud University, Nijmegen, Netherlands, [2]QLVL, Department of Linguistics, KU Leuven, Leuven, Belgium

Dialectometry studies patterns of linguistic variation through correlations between geographic and aggregate measures of linguistic distance. However, aggregating smooths out the role of semantic characteristics, which have been shown to affect the distribution of lexical variants across dialects. Furthermore, although dialectologists have always been well-aware of other variables like population size, isolation and socio-demographic features, these characteristics are generally only included in dialectometric analyses afterwards for further interpretation of the results rather than as explanatory variables. This study showcases linear mixed-effects modelling as a method that is able to incorporate both language-external and language-internal factors as explanatory variables of linguistic variation in the Limburgish dialect continuum in Belgium and the Netherlands. Covering four semantic domains that vary in their degree of basic vs. cultural vocabulary and their degree of standardization, the study models linguistic distances using a combination of external (e.g., geographic distance, separation by water, population size) and internal (semantic density, salience) sources of variation. The results show that both external and internal factors contribute to variation, but that the exact role of each individual factor differs across semantic domains. These findings highlight the need to incorporate language-internal factors in studies on variation, as well as a need for more comprehensive analysis tools to help better understand its patterns.

Keywords: computational sociolinguistics, dialectometry, lexical variation, semantic variation, spatial analysis, mixed-effects regression, limburg

OPEN ACCESS

Edited by:
Dong Nguyen,
The Alan Turing Institute,
United Kingdom

Reviewed by:
Alina Maria Cristea,
University of Bucharest, Romania
Yves Scherrer,
University of Helsinki, Finland
John Nerbonne,
University of Freiburg, Germany

****Correspondence:***
John L. A. Huisman
j.l.huisman@let.ru.nl

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *15 February 2021*
Accepted: *14 May 2021*
Published: *04 June 2021*

Citation:
Huisman JLA, Franco K and van Hout R (2021) Linking Linguistic and Geographic Distance in Four Semantic Domains: Computational Geo-Analyses of Internal and External Factors in a Dialect Continuum. Front. Artif. Intell. 4:668035. doi: 10.3389/frai.2021.668035

INTRODUCTION

Dialect geography deals with the spatial components of human communicative processes or, on a more abstract level, with the relationship between space and social behavior. Social behavior results in spatial patterns of language variation and change. Languages change as speakers accommodate their speech patterns during interactions with their most common conversational partners—their speech community (Bloomfield, 1933)—and for logistical reasons, these interactions occur more frequently and intensely between people that are geographically close to each other. Previous work has shown that linguistic features first spread across communities that share dense interaction, and then expand into the rest of a language area—a process called diffusion (see Gerritsen and van Hout, 2006, for an overview). As a result of this process, the linguistic varieties of neighboring communities generally differ only slightly (Chambers and Trudgill, 1998). In contrast, contact between geographically distant communities tends to be less frequent. Accommodation therefore occurs to a lesser degree, resulting in communities whose linguistic varieties resemble each other less and less the farther apart they are (Heeringa and Nerbonne, 2001). Linguists often call this gradual pattern a dialect continuum, and many have been studied. For example, Nerbonne (2010) investigated six areas (variation across the Bantu languages, in Bulgaria, Germany, across the United States East Coast, and in the Netherlands and Norway) and found that linguistic distance continuously increases over geographic distance, but that the magnitude of this increase diminishes as geographic distances become larger.

Dialectometry aims to objectively measure linguistic relationships between dialects (Séguy, 1971) and the procedure followed in most studies is based on methods used by the Salzburg school of dialectometry, which compare a geographic distance matrix with a linguistic (dis)similarity matrix (see Goebl, 2006, for an overview). These matrices code pairwise geographic distances and linguistic (dis)similarities between all locations in a language area. The online dialectometry tool *Gabmap* (Nerbonne et al., 2011) developed at the University of Groningen follows this same conceptual structure, separating the geographical and linguistic information in two different matrices. To connect the two, Goebl (2006) refers to the Pearson correlation for comparing linguistic and geographical distances. However, as the assumption of independence of observations is violated when dealing with distance values, Mantel (1967) suggested a permutation technique for evaluating significance in such cases. Nerbonne and Heeringa (2007) used this Mantel test to investigate the correlational structure of the linguistic and geographical distances in the Netherlands. However, subsequent reviews of methods in dialectology appear not to mention the Mantel test again (e.g., Wieling and Nerbonne, 2015), despite its wide application in ecology (for overviews, see e.g., Legendre and Legendre, 2012; Zuur, Ieno and Smith, 2007).

A rather pressing issue with such correlational approaches is how to include external factors other than geography in the analysis. Although dialectologists have always been well-aware of the potential influence of language-external variables such as population size, isolation and socio-demographic features, these characteristics are generally only included in dialectometric analyses afterwards. The interpretative maps produced as output of many Salzburg style studies—and also Gabmap—emphasize the value of visualization. Language-external factors are mostly used for further interpretation of the results rather than as explanatory variables in a more formal, statistical model. In addition, one of the downsides of aggregating measures of linguistic distances is that differences between the linguistic variables involved in computing linguistics distance are smoothed out (e.g., Schneider, 1988). This is especially relevant for measures based on lexical variation, as language-internal semantic characteristics have been shown to affect the distribution of lexical dialect data (cf. Speelman and Geeraerts, 2008; Franco et al., 2019b).

One development put forward to address this is the use of generalized additive mixed-effects regression modeling (GAM) as applied by Wieling (2012), in which both geographic and social predictors are included in a regression design. The analysis is done on linguistic distances between a series of observed points (the dialects) and a reference point (the standard language). Explanatory variables that have been studied using this method include both language-external (e.g., community size, speaker education level), and language-internal factors (e.g., word frequency, grammatical category). The strength of the GAM approach is that the models can include random factors, which allows for more precise control over outlying locations and linguistic items or elements. The GAM has been successfully applied to dialectal variation in e.g., Dutch (Ko et al., 2014), Italian (Wieling et al., 2014), and Catalan (Wieling et al., 2018). Wieling and Nerbonne (2015) provide an overview of other quantitative work on multivariate spatial analysis of language variation, e.g., quantitative counterparts to the traditional identification of isoglosses and dialect regions (Grieve et al., 2011).

The GAM approach clearly shows the advantage of regression analysis in explaining linguistic variation, but one particular limitation is the use of a single reference point in defining (dis)similarity. While the analysis provides valuable insights into which factors play a role in differentiation from the standard (the reference point), it is more limited in exposing overall patterns of variation that exist in the dialect area as a whole, i.e., how variation between non-standard varieties is patterned. In addition, when dialect areas are spread across multiple countries—such as Limburgish, the area under investigation here—it is even harder to determine what to use as a reference point. In dialectometry however, we prefer to compare each location to all other locations. This gets rid of the need for a single point of reference, and it helps understand the patterns of linguistic variation across the entire linguistic area. This challenge was recognized by Wieling and Nerbonne (2015), who advocated the search for an approach that is able to incorporate information from the linguistic landscape as a whole, while at the same time including non-linguistic factors as explanatory variables. In this paper, we showcase two regression methods that keep all location-by-location comparisons intact.

As a first step, we use Multiple Regression on distance Matrices (MRM; Lichstein, 2007). MRM is an extension of the (Partial) Mantel test on two (or more) distance matrices. In essence, the relationship between the Mantel test, the partial Mantel test, and MRM is the same as between analyses of correlations, partial correlations and multiple regression. The main advantage of MRM over the Mantel test is that while the Partial Mantel test combines the different explanatory factors into a single distance matrix, MRM allows for each factor to be included individually, which makes it possible to assess their individual importance. As such, it is more flexible in terms of the

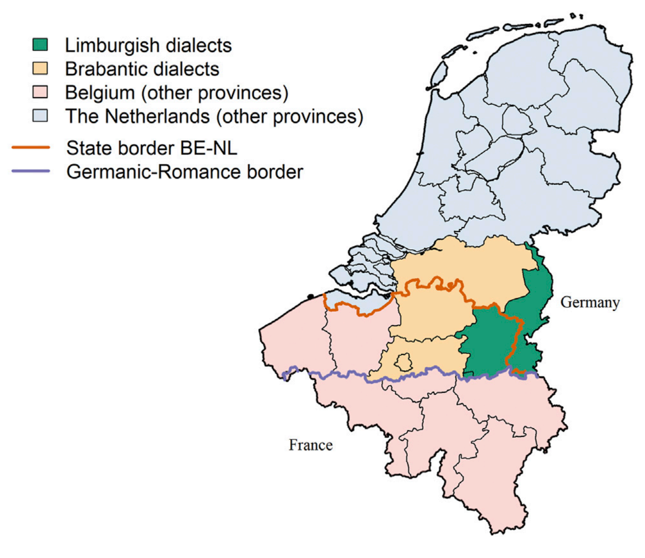


FIGURE 1 | The Limburgish dialect region (in green). The brown line within the green area is the boundary between the Dutch (on the **right**) and the Belgian (on the **left**) province of Limburg. The purple line in the south of the dialect region shows the boundary between the Dutch-speaking (northern) and French-speaking (southern) part of Belgium. The other provinces of the Netherlands and Belgium are colored light-blue and pink.

types of data that may be analyzed (e.g., binary, continuous), and it provides estimations of explained variance. In addition, significance testing is done on the basis of random permutations to avoid overestimating the significance of the correlations. In contrast to earlier studies using GAM with a single point of reference (e.g., Wieling, 2012; Wieling et al., 2014; Wieling et al., 2018), the use of distance matrices allows to include the distances between all locations in a single analysis.

We use the MRM analysis as an in-between step toward a more comprehensive analysis based on linear mixed-effects modeling (LMER). LMER shares with GAM the potential to include random variables, and was applied by Huisman, Majid and van Hout (2019) to analyze the role of external factors in linguistic variation across Japan. Here, we expand on this by applying the method to a much larger dialect database for the Limburgish dialect continuum in the Netherlands and Belgium, and by including language-internal factors. Crucially, LMER renders the use of the location-by-location matrix format obsolete, which is particularly important for variables that cannot be coded into distance matrices, i.e., language-internal factors such as semantic density and salience. We performed a simulation study to show the strength of linear mixed-effects models (LMER) in comparison to Multiple Regression on distance Matrices (MRM), which we describe in the **Supplementary Material**.

The main aim of this contribution is to chart the many promising possibilities of our application of linear mixed-effects regression modelling (LMER) on full pairwise distance matrices to simultaneously investigate external and internal drivers of language variation. By incorporating techniques from quantitative ecology into the dialectometric methodology, we can develop strong explanatory models of patterns of language variation on a large spatial scale. Following dialect geography, which deals with the intertwining of linguistic and geographical variation, the primary link in explaining differences in a dialect continuum is geographic distance, an external factor. However, we will show that geographic distance is the entrance to test the role of additional factors, internal or external. Our implementation of LMER provides a more comprehensive analysis tool that offers supplementary techniques to analyze and understand patterns of geographical language variation, that seem to be superior in several respects to techniques currently used in dialectometry.

MATERIALS AND METHODS

The Limburgish Dialect Continuum

The lexical data analyzed in this paper comes from the Limburgish dialects, spoken in the northeast of Belgium and

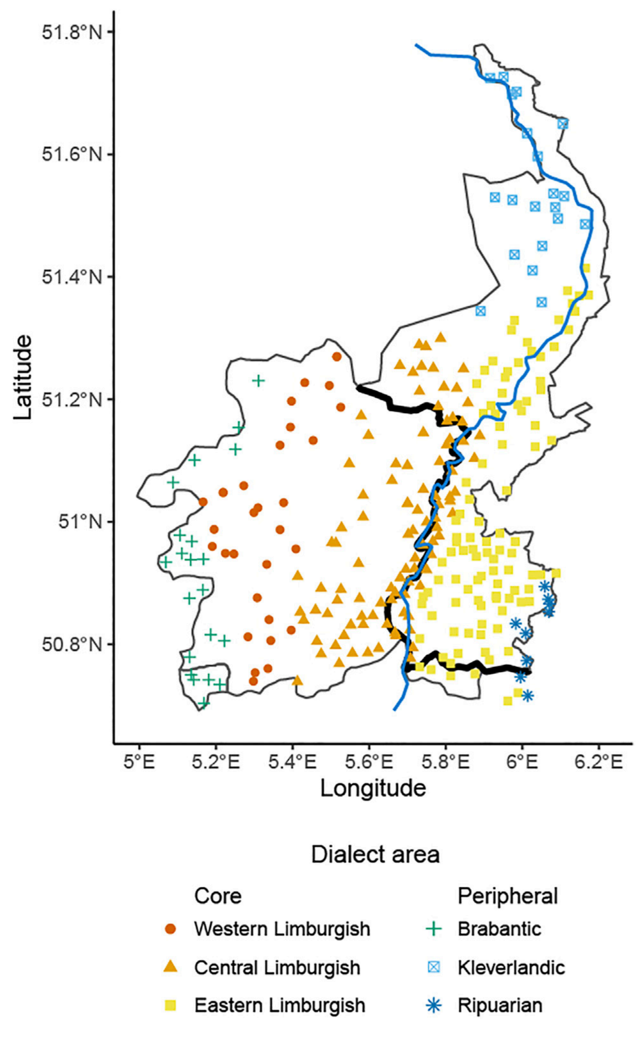


FIGURE 2 | Map of locations included in the database, with their classification into one of six dialect areas.

in the southeast of the Netherlands (marked in green on **Figure 1**). In the south, the dialect area is demarcated from the Romance language area by the Germanic-Romance border (marked in purple). In the east, the dialect area is demarcated by the national border with Germany. The German and Dutch dialects (which include the Limburgish dialects) historically form a dialect continuum—some of the dialects spoken in the south of Limburg (e.g., the Ripuarian dialects, see **Figure 2**) in the Netherlands can even be considered dialects of German as they underwent the second Germanic consonant shift. In the north and west, the dialect area borders the Brabantic dialect area, another dialect area of Dutch (marked in orange on **Figure 1**). Although there is some discussion about where the Limburgish dialects end and the Brabantic dialects begin, the demarcation is often equated with the provincial borders in the Netherlands and in Belgium (Weijnen, Goossens and Goossens, 1983). Thus, it is accepted that Limburgish dialects are spoken in the Belgian province of Limburg and in the province of Limburg in the Netherlands, whereas Brabantic dialects are used in the Belgian provinces of Flemish Brabant (including Brussels) and Antwerp, and in the province of North Brabant in the Netherlands.

Six subregions can be distinguished within the Limburgish dialect area: Western Limburgish, Central Limburgish, Eastern Limburgish, Kleverlands, Ripuarian, and Brabantic (Van de Wijngaard and Keulen, 2007). The latter three areas are peripheral and transitional areas that share a border with other dialect regions. In addition, the national border between Belgium and the Netherlands runs through the dialect area. Its current position was officially determined in 1839 when the independence of the nation of Belgium was definitively recognized internationally.

The Dictionary of the Limburgish Dialects

The linguistic data we used come from the *Dictionary of the Limburgish Dialects*. These data offer a firm basis to gain insight into patterns and processes of linguistic variation and change. They cover a large part of the lexicon, comprising thousands of concepts belonging to all aspects of human life. Importantly, the data were collected at the concept level, avoiding possible bias in the data selection process as the researcher does not need to determine which variants are synonymous across locations. The data were collected highly systematically, mainly through large-scale dialect questionnaires distributed between 1960 and 1990, in which lexical variants were elicited for every concept. In addition, the paper version of the dictionary contains data from additional sources (e.g., local dialect dictionaries and specialized terminological dialect collections), which was used to complement the dictionary entries.

For our analyses, however, we only used the data that were collected by means of the questionnaires. Furthermore, we used the digitized version of the dictionary, which has in recent years also become available online (http://www.e-wld.nl), as the paper dictionary does not contain all the questionnaire data as a result of editorial work. For example, concepts without any variation across the entire dialect area, or questions that produced messy data as respondents found them difficult, were not included in the paper dictionary. However, for our analysis, we use all the data available in the database, including concepts without variation or with messy responses.

The Four Semantic Domains

The dictionary is divided into three large parts, covering all aspects of human life in the first half of the 20th century: agrarian terminology, non-agrarian professional terminology, and general vocabulary. Every part consists of about a dozen volumes, each containing the vocabulary for one specific semantic domain. In the analyses presented here, we focus on four semantic domains from the general vocabulary: *Church and religion*, *Clothing and personal hygiene*, *the Human body*, and *Society and education*. These domains were selected because the concepts they represent vary along two axes. First, they differ in the degree to which they contain basic vocabulary concepts (*the Human body*) versus culturally variable concepts (*Church and religion*, *Clothing and personal hygiene*, and *Society and education*). Second, the semantic domains with cultural vocabulary show differing degrees of top-down standardization. For example, the *Church and religion* field is characterized by a high degree of standardized vocabulary, often of Latin origin, related to general religious practices and traditions. In contrast, there is no high degree of standardization for the *Clothing and personal hygiene* field. In addition, the *Church and religion* domain contains concepts relating to the Catholic church in particular, but as will be explained below, the Catholic religion does not play an equally large role throughout the dialect area.

Table 1 provides an overview of the data per domain in the database. In most semantic domains, data is available for 175 or more locations in the Limburgish dialect area, but this number is lower in *Church and religion* (114), where the distribution of locations with data is less dense across the dialect area (see **Figure 3**, where locations without data per semantic field are shown with a grey circle, whereas locations with data are marked in red). In addition, **Table 1** also presents the information of the language-internal factors that will be included in the analysis below. The domains and these language-internal factors are discussed in more detail in the following sections.

The *Church and Religion* Domain

The *Church and religion* domain contains concepts relating to different aspects of Catholicism. It consists of five subsections: *In and around the church building*, *Liturgy and devotion*, *Catholic Holy Days and rites*, *Catholic belief and faith*, and *The clergy*. **Table 2** contains example concepts for each subsection. The first subsection, *In and around the church building*, consists of concepts relating to the interior and exterior of a typical church building in the Low Countries—e.g., the names for the typical parts of a church, such as the baptistery, the sacristy, the church tower and bells and the cemetery that is typically found around a Catholic church. The second subsection, *Liturgy and*

TABLE 1 | Data per semantic domain in the database.

Semantic domain	Number of locations	Number of concepts	Number of subsections per level of depth			Ratio of multi-word concepts	Concept length	
			1	2	Max		Mean	Median
Church and religion	114	592	5	19	35	0.30	12.9	11
Clothing and personal hygiene	188	323	5	18	43	0.48	13.8	12
Human body	179	180	3	14	18	0.32	10.7	9
Society and education	175	462	4	19	55	0.21	9.9	9

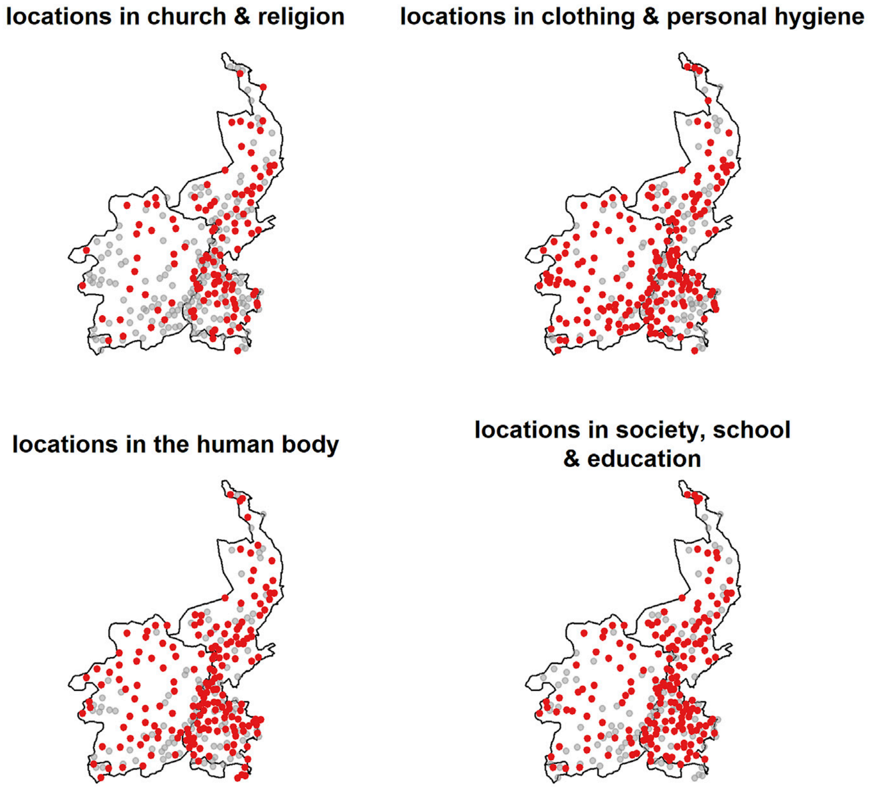

locations in church & religion
locations in clothing & personal hygiene
locations in the human body
locations in society, school & education

FIGURE 3 | Locations available per semantic domain (in red). Grey circles indicate that no data is available for a particular semantic field.

TABLE 2 | Examples of concepts in the subsections of the *Church and religion* domain.

Subsection	Examples
In and around the church building	Church, leaded window, credence (table), church bell, tombstone
Liturgy and devotion	Early mass, offertory, holy water, rosary, to pray
Catholic holy days and rites	Patron saint, advent, good friday, confirmation, confession
Catholic belief and faith	Catechism, devil, baby jesus, fasting
The clergy	Pope, franciscan, monk, dean

devotion, mostly contains concepts relating to the Catholic mass, such as its different types (e.g., in the morning, at night, for children), its typical parts, and different prayers (e.g., the Lord's Prayer, Hail Mary). The third subsection, *Catholic Holy Days and rites*, contains concepts relating to the Catholic Holy Days and the Catholic Calendar (e.g., Christmas, Easter), and also describes the seven sacraments (i.e., baptism, marriage, confession, etc.), as well as the Catholic funeral. The fourth subsection, *Catholic belief and faith*, contains more general aspects of Catholic faith, such as religious concepts (e.g., purgatory, fallen angel, miracle), as well as

TABLE 3 | Examples of concepts in the subsections of *Clothing and personal hygiene* domain.

Subsection	Examples
Clothing	Smock, undervest, to fit, women's coat
Headgear	Beret, hat, pom-pom of a bonnet, bowler hat
Foot- and legwear	Barefoot, women's shoe with medium or high heel, clog, sock
Jewelry and ornaments	Watch, medallion, jewelry, sequin
Personal hygiene	To shower, to brush teeth, toothpick, razor

TABLE 4 | Examples of concepts in the subsections of the *Human body* domain.

Subsection	Examples
The body and the body parts	Short, curly hair, eye, navel
Organs and their functions	To breathe, stomach, kidney, diaphragm
The senses	To see, to wink, flavor, to listen attentively

different virtues and sins. The final subsection, *The clergy*, contains the different names for people belonging to the clergy.

Previous research on this semantic domain in the *Dictionary of the Limburgish Dialects* and its counterpart for the Brabantic dialects has shown that there is a large number of loanwords from Latin, traditionally an important language of the Catholic church (Franco et al., 2019a). The frequency distribution of Latin loanwords across the two dialect areas was largely similar, with only minor differences between Brabant and Limburg. This indicates that these loanwords were not distributed across the dialect areas by linguistic diffusion, as this would have resulted in a more wave-like pattern of their distribution. Instead, the loanwords were likely introduced in all dialects as necessary borrowings for religious concepts and then transmitted from generation to generation (cf. Labov, 2007). These findings are not surprising as the Catholic religion has held a strong position in the Low Countries—especially in the south of the language area, which includes the Limburgish and Brabantic dialects. For example, data from Schmeets (2014:6) indicates that at the beginning of the 20th century (in 1909), 100% of the people living the Dutch province of Limburg self-reported as being Catholic. In1987,[1] close to the time when most of the data for this semantic domain was collected, this number had only dropped to 89%. Due to the fact that the rites and structure of the Catholic church are standardized, we may also expect that the effect of geographic distance in this semantic domain is smaller than in other semantic domains.

The *Clothing and Personal Hygiene* Domain

The *Clothing and personal hygiene* domain consists of five subsections: *Clothing*, *Headgear*, *Foot- and legwear*, *Jewelry and ornaments*, and *Personal hygiene*. Example concepts for each subsection are shown in **Table 3**. Most concepts belong to the first subsection, *Clothing*. The semantic domain as a whole contains culturally variable vocabulary, as evidenced, for instance, by the fact that some of the concepts have fallen out of use more recently (e.g., jerkin, nightcap). Concepts related to clothing have been shown to be prone to lexical borrowing (Tadmor, 2009) and previous work has confirmed that this is also the case in the Limburgish data. Many loanwords from French are in use, especially in the Belgian part of the dialect area (Franco et al., 2019a). We may therefore expect that this domain is prone to patterns of diffusion (Labov, 2007), resulting in larger linguistic differences between locations that are further away from each other.

The *Human body* Domain

The *Human body* domain consists of three main parts: *The body and its parts*, *Organs and their functions*, and *The senses*. Example concepts for each of these parts are provided in **Table 4**. Many body part concepts have been included in basic vocabulary lists. For instance, 19 concepts on the 100-item Swadesh list (Swadesh 1955) and 25 concepts on the 100-item Leipzig-Jakarta list (Tadmor, 2009) are part of this domain. However, the dictionary also contains entries for several jocular terms for body parts (e.g., for the head or mouth), children's names for body parts, taboo meanings (e.g., names for male and female genitalia), as well as concepts that are cognitively less salient, referring to parts of the body that might not turn up often in everyday conversations (e.g., dimples, or the upper part of the back). For the basic vocabulary concepts, we expect to find little lexical variation, whereas more variation can be expected across the dialect area for the jocular terms, children's names, and taboo and non-salient concepts (see Speelman and Geeraerts, 2008; Geeraerts and Speelman, 2010; Franco et al., 2019b).

The *Society and Education* Domain

The *Society and education* domain consists of four diverse subsections relating to the different aspects of public life: *People and society*, *Societal organization*, *Transport*, and *School and education*. **Table 5** shows examples for each subsection. The first subsection, *People and society*, is the largest and comprises topics on social life in the community (e.g., names for neighbors and visitors, going to parties, and friendship and animosity), trade (e.g., buying and selling, names for monetary units, names for commercial places, names for property), social etiquette, as well as language and communication. The second subsection, *Societal organization*, describes topics such as the organization of the state, policing, the judiciary, and war. The third and fourth subsections, *Transport* and *School and education*, are more limited in size, containing concepts relating to different modes of transportation (e.g., by air, water or road), and the organization of the education system. As that many concepts in this broad domain are related to the way a state is organized, we may expect that state-level decisions (e.g., the monetary unit that is used in a particular country) lead to small differences between locations

[1] The most important questionnaire used for this semantic field was distributed across the dialect area in 1989.

TABLE 5 | Examples of concepts in the subsections of the *Society and education* domain.

Subsection	Examples
Man and society	Company, to peddle, night owl, market booth, (Dutch) guilder, impolite, fairy tale, to complain
Societal organization	Mayor, liberal, charge, perjury, soldier, to fall in battle
Transportation	Pedestrian, women's bicycle, train, steamboat, airplane, to travel
School and education	Boarding school, teacher, ruler, report card

even when they are geographically far away. However, larger differences between locations may be found for concepts that describe aspects of societal organization at a lower level (e.g., neighbors, staying out late). Finally, as the Limburgish dialect area spans two countries, the national border might play a large role in variation for this domain.

Explanatory Factors

In the analyses, we examine the effect of both internal and external factors on linguistic distance between locations. We selected five language-external and two language-internal factors that have been previously used in dialectology and related fields to predict linguistic variation and change, each of which is described in detail below.

Language-External Factors

Geographic Distance

The first factor we investigated is geographic distance. This factor has a well-known effect on linguistic distances: the further two locations are located from each other, the more their language is expected to be different (e.g., Bloomfield, 1933; Chambers and Trudgill, 1998). Geographically, this results in linguistic structures being distributed across a dialect area in a wave-like pattern. An explanation for this finding is the frequency of contact between language users—the principle of density of communication (see Bloomfield, 1933).

For the analyses, we determined latitude and longitude coordinates for all locations in our dataset and calculated straight line geographic distance in kilometers between every pair of locations. As logarithmic distance has been shown to predict linguistic differences more accurately in several studies (e.g., Nerbonne and Heeringa, 2001), we also calculated the natural logarithm of the geographic distances.

Separation by Water

Another factor that has been argued to affect linguistic distance is isolation: the more isolated a speech community is (e.g., due to natural borders such as rivers or mountain ranges), the less similar their language will be to other related surrounding varieties. Water as a natural border has been shown to influence variation in the Dutch language area. For instance, the word for "purse" differs across the islands of the Dutch province of Zeeland: *borre* is used on the island of Goeree, *bozze* on Schouwen and Southern Beveland and *beuze* on Walcheren and Noord-Beverland (Weijnen, 1966). This effect of separation by water has also been shown across the Japanese archipelago, for both varieties of Japonic (Lee and Hasegawa, 2014; Huisman et al., 2019) and Ainu (Lee and Hasegawa, 2014).

Through the Limburgish dialect area runs the river Meuse, which partly forms the border between Belgium and the Netherlands. In contrast to oceanic barriers discussed above, the role of rivers in creating isoglosses in dialect areas is less clear. A river can both separate and unite depending on its navigability (Weijnen, 1966), which is why it is interesting to include the Meuse in the current study. For the analyses, we first determined, for each location, whether it is located to the west or to the east of the river and then coded, for each pair of locations, whether they are located on the same side of the Meuse (coded as 0) or on opposite sides (coded as 1).

Population Size

Another factor that has been shown to influence linguistic distances is population size. In Trudgill's gravity model, for example, the effect of geographic distance is mediated by population size: language changes are expected to first diffuse from one large city to another, and to only be adopted later in smaller locations (Trudgill, 1974; Chambers and Trudgill, 1998). The explanation is again contact between speakers: the language used in large urban centers influences the language of smaller locations.

Because the locations available in the *Dictionary of the Limburgish Dialects* are often neighborhoods or districts that are part of a larger administrative community, obtaining accurate and specific population sizes is not without challenge. Especially in Belgium, figures are only publicly available at the administrative community-level. In addition, the questionnaires for the dictionary were distributed several decades ago and obtaining public *historical* population data for every location in the database is even more difficult. Finally, many administrative communities were merged in Belgium in the last quarter of the 20th century (i.e., after the dictionary project started; see De Ceuninck, 2009). For example, Heusden and Zolder in the center of Belgian Limburg were historically two separate administrative communities but merged into a single administrative community (Heusden-Zolder) in 1977.

To handle these challenges, we opted for a systematic procedure that ensures that the population size data we collect is as comparable as possible across the two countries and the dialect area as a whole.[2] For most of the data, we can rely on census data collected by the national governments. The oldest data available in Belgium stems from 2008 (Stat Bel, 2021), and so we also used data from 2008 in the Netherlands (CBS, 2021). While the Dutch data also provides population sizes for districts and neighborhoods *within* an administrative community (CBS, 2017), we have not found the same type of information for the locations in Belgium. For locations for which data was not

[2] The data was collected in March 2019.

directly available, we used information from Wikipedia.[3] Finally, for 8 locations (1.2% of the total) we did not find any figures for their population size. The median population size of all locations is equal to 2,242. Since the locations without population sizes are generally small, we rounded this number down to 2,000.

As the techniques we used require distance matrices as their input (see *Statistical Analysis*), we calculated the (absolute) *difference* in population size between all pairs of locations in the database. In addition, to account for magnitude effects, we performed log transformation on these differences. We decided to include population size as an individual predictor rather than a gravity-based approach because this has the potential to determine the strength of population size on its own—which is important given that previous work has shown that in some cases the bulk of the influence exerted by gravity comes from distance alone, e.g., Nerbonne and Heeringa (2007).

National Border

The third factor we investigated is the national border between Belgium and the Netherlands. The effect of national borders is well-known in the dialectological literature and may lead to dialect divergence due to convergence with the national language (Hinskens et al., 2000). Dutch is a pluricentric language, with Netherlandic and Flemish standard varieties (Willemyns, 2013). It has been shown that the border affects the dialect variants used (Cajot, 1977; Gerritsen, 1999; Franco et al., 2019a).

For the analyses, we first determined, for each location, whether it is located in Belgium or the Netherlands, and then coded, for each pair of locations, whether they are located in the same country (coded as 0), or in different countries (coded as 1).

Dialect Area

Finally, to control for potential effects of increased uniformity within dialect areas (see e.g., Shackleton, 2005; Nerbonne, 2013 for previous uses of dialect area as explanatory factor), we classified each location into one of six subgroups based on the classification from Van den Wijngaard and Keulen (2007), as outlined above: the three core areas Western Limburgish, Central Limburgish, and Eastern Limburgish, and the three peripheral areas Brabantic, Kleverlandic, and Ripuarian—see **Supplementary Table S3** for the classification of each location. For the analyses, we then coded, for each pair of locations, whether they are from the same dialect area (coded as 0), or from different areas (coded as 1).

Language-Internal Factors

Semantic Density

Semantic density concerns the extent to which a semantic domain is carved up into lexicalized concepts by language users. Some semantic domains are very dense with many different meanings being lexicalized, whereas other domains use semantically broader and vaguer concepts. Semantic density can differ across languages. For instance, Majid and Burenhult (2014) showed that very few olfactory concepts are lexicalized by speakers of English, but that speakers of Jahai, nomadic hunter-gatherers of the Malay Peninsula, can name and distinguish smell as easily as color, as smell takes up a prominent place in their everyday life and communication. While this shows that, at least at the broader culture level, semantic domains can differ in density due to differences in communicative needs (e.g., Kemp et al., 2018), little research exists on such differences between domains within a dialect area. We therefore included this variable in the analyses to test whether semantic density correlates with linguistic distance. Following the results of previous work, we presume that increased cultural relevance leads to increased semantic density, which in turn leads to decreased lexical variation.

For the analyses, we used two approaches to represent semantic density (see **Table 1**). First, we determined, for each domain, the number of subdomains into which each respective semantic domain is divided, by using the subsections identified in the dictionary. We selected three levels of granularity: the first (broadest) level of subsections, the second level of subsections, and the deepest level (ranging between 4 and 6, depending on the domain). Secondly, we determined the total number of concepts in each domain, and calculated the average number of concepts per subsection at each level of granularity.

Salience

Salience concerns the extent to which a particular meaning is familiar to language users. For example, concepts like pants, shoe or shirt are highly salient: most human beings in industrialized societies are probably familiar with them and come into contact with them every day of their life. In contrast, concepts like jerkin or bowler hat are much less salient (at least nowadays) because these concepts represent objects that people no longer make use of often.

The notion of salience was introduced in Geeraerts et al. (1994), who relate it to the basic-level hypothesis (Berlin, 1972, Berlin, 1978; Berlin et al., 1973). This hypothesis is based on the fact that, cross-linguistically, folk biological classifications consist of a limited set of taxonomical levels, which reflect the degree of salience of the organisms involved. Referents with a high degree of salience (e.g., oak, robin), constitute the core of any folk biological organization and, thus, the basic level: "[a]t this rank, both plants and animals appear perceptually most distinct to the human classifier, and these differences in morphology and behavior virtually 'cry out to be named'" (Berlin, 1978: 24). Properties of categories at the basic level are that their name is highly frequent (Rosch et al., 1976) and typically consists of a short, primary lexeme, i.e., a non-compositional simplex word like *oak* or *robin* (Berlin, 1972: 54).

Geeraerts et al. (1994) showed that the concept of the basic level is problematic when applied to other types of categories, like

[3] Available at https://www.wikipedia.org/. Wikipedia is of course not an ideal source for this type of information, but we believe there is merit in this case because the pages devoted to neighborhoods or districts within an administrative community often display a large degree of local pride identity, e.g., by naming famous people from the community or by describing its history and local traditions. This makes it unlikely that incorrect information will be added to the website. In addition, every Wikipedia article also has a "Talk page", which is used as a discussion board to improve the article (e.g., by correcting faulty information). In total, for 155 out of 660 communities, Wikipedia data was used. The Wikipedia data is mostly based on domicile information and typically describes the number of residents for a particular year, most often between 2006 and 2018.

artefacts such as clothing. First, the hypothesis presupposes a neat taxonomical organization of the lexicon, because it is based on inclusion relationships. However, a clothing item like *broekrok* "culottes, lit. pants-skirt" poses a problem in this view as it is difficult to place in a taxonomy in which skirts and pants form the basic level. The authors argue that the lexicon is organized in the form of overlapping taxonomies that are all based on different dimensions. Secondly, and more importantly, Geeraerts and colleagues show that for artefacts like clothing items, onomasiological typicality exists between categories *on the same level* of a taxonomical hierarchy as well. For this reason, they propose to take into account a generalized notion of onomasiological salience, which they relate to Langacker (1987) notion of entrenchment. Crucially, this approach allows them to show that differences in salience, both between and within taxonomical levels, correlate with naming preferences, including the fact that concepts that are more entrenched are more likely to be named with simplex forms. Later studies have shown that concepts with a higher degree of salience not only correlate with naming preferences, but also with decreased dialectal variation (Speelman and Geeraerts, 2008; Geeraerts and Speelman, 2010; Franco et al., 2019b). For example, in the semantic field of the human body, the Limburgish dialect dictionary only contains a single word for a highly salient concept like blood, whereas a lot more variation occurs for less salient concepts like the little dents between the knuckles, or bristly (w.r.t. hair on one's head).

For the analyses, we used two aggregate measures to determine the average degree of salience per domain, which we based on the concept names available in the dataset (see **Table 1**). These were usually the prototypical Standard Dutch word for that concept. First, following the tendency for high salience concepts to be named with non-compositional simplex words, we calculated, per domain, the ratio of multi-word concepts compared to the total number of concepts. Secondly, following the correlation between salience and word length that has been described, we calculated the mean and median concept length in number of characters [a similar approach was used in Franco et al. (2019b), Speelman and Geeraerts (2008), and Geeraerts and Speelman (2010)]. We expect that linguistic distances will be larger in semantic domains with more multi-word concepts or with longer concepts.

Statistical Analysis

Linguistic Distances

Several measures of linguistic distance have been used in dialectometry, such as the binary same vs. different coding, or Levenshtein distance, to calculate how much two forms differ from each other (see Nerbonne and Kleiweg, 2007 for a discussion). The data we extracted from the *Dictionary of the Limburgish Dialects* uses a standardized coding system based on cognacy, which the editors of the dictionary invested based on their expertise in dialectology and historical linguistics (van Hout et al., 2014). A major advantage of coding entries on the lexical level was that it was no longer necessary to make (sometimes rather arbitrary) decisions about the level of phonetic detail to be coded. This is especially relevant because in this case, the volunteers filling in the many dialect questionnaires were not linguistic professionals. As a result, specific surface forms are collapsed into a general entry. For example, *toŋ* (Maastricht dialect), *tuŋ* (Hasselt dialect) and *tsoŋ* (Kerkrade dialect) for "tongue" are all in the dictionary as TONG. This has consequences for measuring distances. The finer phonetic details are not transcribed in the standardized forms. This is likely why preliminary results showed that, based on measures of explained variance and residual scores, string edit distance algorithms were outperformed in our data by methods for binary distance coding.

As the quality of these dictionaries is thus found at the *lexical* level, we used a measure based on the Weighted Identity Value (*Gewichteter Identitätswert*, GIW; Goebl, 1984), which codes binary differences, but takes into account how frequent particular word forms are and weighs them accordingly. We used Gabmap (Nerbonne et al., 2011) to calculate linguistic distances between all pairs of locations based on Gabmap's use of $d = 1\text{-}GIW$—a measure we will call Weighted Dissimilarity Value for the remainder of this paper—for each semantic domain separately.

For the regression analyses, we also computed the logit value of the distance measure to tackle two problems: 1) the range of the dependent variable, 2) the non-linear relation between this variable and the natural logarithm of geographic distance.

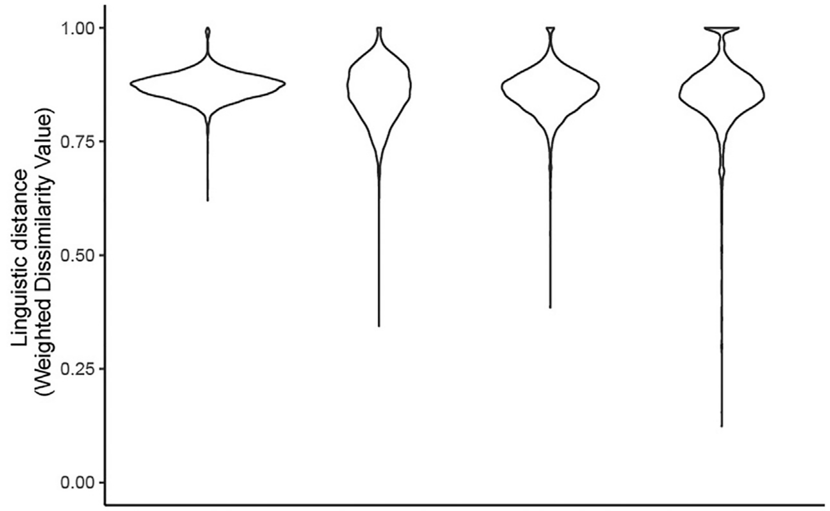
Correlational Analyses

To assess how patterns of linguistic variation can be explained by language-external factors, we used Mantel correlations, which are widely used in ecology to analyze relations between measures coded in pairwise distance matrices. The analyses were performed in *R*, using the *ecodist* package (Goslee and Urban, 2007). We used the *mantel* function to calculate partial Mantel correlations (using 10,000 permutations and 1,000 bootstrap iterations on 95% confidence intervals) between linguistic distances and each of the language-external factors—for each semantic domain separately. In addition, we used the *MRM* function to perform Multiple Regression over Distance Matrices (using 10,000 permutations) for each domain separately.

Regression Analyses

The need for data coded as distances matrices required for analyses based on Mantel correlations brings about several shortcomings. One factor that cannot be addressed using such correlation analyses is the inherent uniqueness that individual varieties included in this (or any linguistic) study all possess. We believe that taking this individual variability of dialects into account allows for better estimation of the contribution of explanatory variables (both external and internal), which is why we used linear mixed-effect regression (LMER) modelling to repeat the MRM analyses, adding to this each individual location as a random factor to account for their inherent uniqueness.[4]

[4] The models that will be discussed below, include locations as random intercepts. Although we also built models with random slopes, these models did not converge.



	Church and religion	Clothing and personal hygiene	Human body	Society and education
Minimum	.620	.344	.385	.123
Maximum	1.00	1.00	1.00	1.00
Median	.872	.863	.862	.858
Mean	.871	.858	.858	.856
SD	.028	.060	.048	.073

Semantic domain

FIGURE 4 | Distribution of linguistic distances across four semantic domains.

In addition, the language-internal factors we are examining in the current study (see above) do not contain information specifically about individual dialects. This means they cannot be coded into distance matrices either and require the use of other methods as well. Again, linear-mixed effects modelling can incorporate such variables, and as such we performed a series of models with the language-internal factors included to assess their contribution to patterns of linguistic variation.

We performed the analyses in R, using the *lme4* package for the modelling, the *reghelper* package to calculate standardized coefficients, the *lmerTest* package for estimates of *p*-values, and the *piecewiseSEM* package to derive pseudo-R^2 values.

RESULTS

Overview

Figure 4 is a violin plot of linguistic distances (as measured through the Weighted Dissimilarity Value) for all unique pairwise location-by-location comparisons across the four domains—excluding comparisons with the same location. The figure shows that even though the mean Weighted Dissimilarity Value was virtually the same across the four domains (0.86–0.87), the range and distribution of linguistic distances differs considerably between them, indicating the value of conducting analyses on a domain basis. The scores are fairly high, indicating a high degree of lexical variability, which may in part be due to the many multi-word answers to the questions in the dialect questionnaires. These multi-word responses can come about in several ways. Sometimes the question that is asked to elicit dialectal responses will by nature elicit a multi-word form. For example, for "a note of 100 franc" multi-word responses, such as *biljet van honderd, briefje van honderd* or *bankje van honderd frank*, are elicited. Similarly, the question "to change your mind" occurs many times with the reflexive pronoun *zich* (specifically in the construction *zich bedenken*). In these cases, the linguistic distances between the words will be small, as identical multi-word responses are also aligned through the Gabmap algorithm. However, in other cases, these multi-word expressions are a sign that the language user is not familiar with the dialect word for the concept in their dialect, and uses a more descriptive response. For instance, for "to grin," one respondent from Susteren used the description *uitgestreken gezicht* (lit. "a straight face"). If many respondents use a large set of these types of descriptive multi-word responses, the linguistic distances will be very large. The effect of these types of multi-word responses on dialect variability is discussed further in Franco et al. (2019b).

Figure 5 illustrates the overall relationship between geographic distance and linguistic distance across the four domains. We plotted, for each location, a LOESS smooth of linguistic distances as a function of geographic distance and included the smooths for all locations into a single plot per

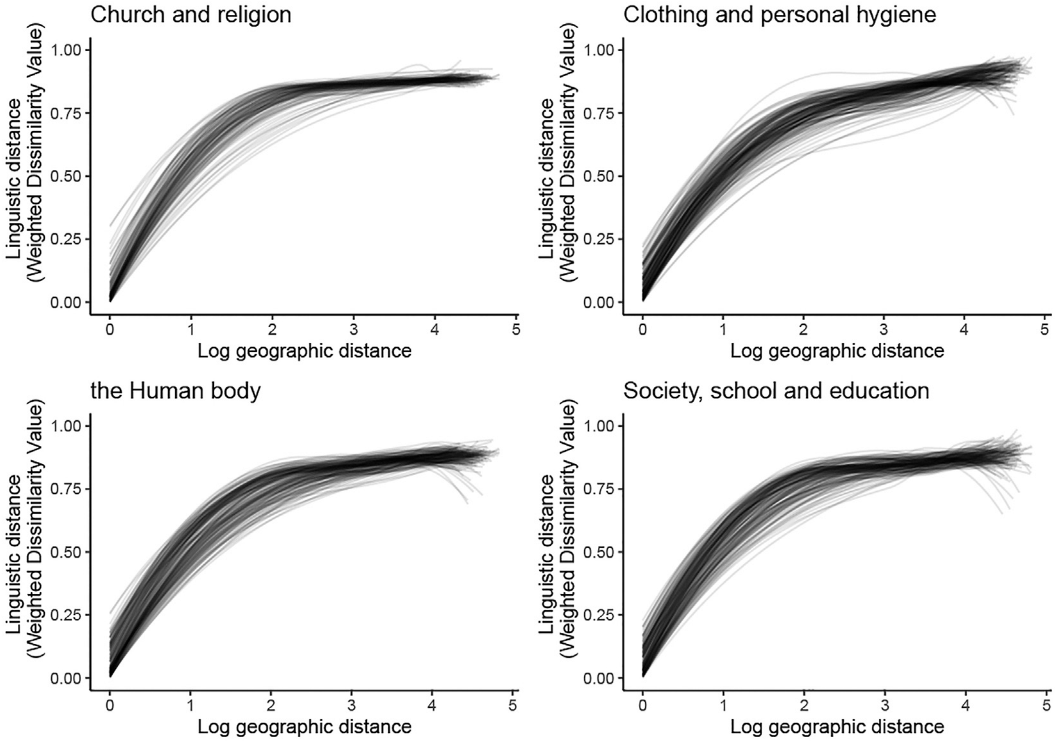


FIGURE 5 | Linguistic distance over geographic distance across four semantic domains.

TABLE 6 | Multiple regression over distance matrices (MRM) results for the *Church and religion* domain.

	Estimate	*P*
Intercept	0.820	<0.001
Log geographic distance	0.010	<0.001
Dialect area	0.003	0.009
National border	0.033	<0.001
Border * distance	–0.003	0.110
Separation by water	–0.023	0.001
Water * distance	0.008	<0.001
Log population difference	0.001	0.266

R^2 = 0.329.

semantic domain, including the comparison of a location with itself. As the figure shows, the overall trend is similar across the four domains, but there are some differences between the individual smooths, indicating that the relationship between geographic distance and linguistic distance differs across both locations and domains. For example, the curves show that lexical dissimilarity rapidly increases over distance in the beginning, but this increase is more pronounced for the *Church and religion* than for the *Society and education* domain. In addition, where linguistic differences in the *Church and religion* domain appear to level off, they keep slightly increasing in the *Clothing and personal hygiene* domain.

Multiple Regression Over Distance Matrices

The first step of our analyses aimed to uncover how linguistic distance is influenced by several language-external factors: geographic distance, dialect area, the national border, separation by water, and differences in population size. The domain-based results are discussed below first, after which we present a summary of the findings across the four domains. As we focused on comparing MRM and LMER as analysis techniques, the discussion below only includes the results of the MRM analyses. However, we have included the Partial Mantel correlations—which showed exactly the same patterns—as well as the correlations between the predictor variables (which were all small to moderate, all r's < 0.5) in our **Supplementary Materials**.

The *Church and Religion* Domain

Table 6 shows the MRM results for the *Church and religion* domain. Main factors that were correlated with linguistic distance were geographic distance, dialect area, the national border, and separation by water. In addition, there was a significant correlation with the interaction between separation by water and geographic distance. The language-external factors accounted for approximately 33% of the variation.

The *Clothing and Personal Hygiene* Domain

Table 7 shows the MRM results for the *Clothing and personal hygiene* domain. Main factors that were correlated with linguistic distance

TABLE 7 | Multiple regression over distance matrices (MRM) results for the *Clothing and personal hygiene* domain.

	Estimate	p
Intercept	0.698	<0.001
Log geographic distance	0.039	<0.001
Dialect area	0.004	0.004
National border	0.071	<0.001
Border * distance	−0.006	0.010
Separation by water	−0.022	0.003
Water * distance	0.005	0.015
Log population difference	0.000	0.987

R^2 = 0.453.

TABLE 9 | Multiple regression over distance matrices (MRM) results for the *Society and education* domain.

	Estimate	p
Intercept	0.795	0.997
Log geographic distance	0.016	0.000
Dialect area	0.002	0.407
National border	0.011	0.374
Border * distance	0.004	0.197
Separation by water	−0.008	0.423
Water * distance	0.003	0.216
Log population difference	−0.001	0.263

R^2 = 0.087.

TABLE 8 | Multiple regression over distance matrices (MRM) results for the *Human body* domain.

	Estimate	p
Intercept	0.769	<0.001
Log geographic distance	0.021	<0.001
Dialect area	0.004	0.006
National border	0.032	<0.001
Border * distance	−0.002	0.510
Separation by water	0.000	0.957
Water * distance	0.003	0.178
Log population difference	−0.001	0.393

R^2 = 0.264.

were geographic distance, dialect area, the national border, and separation by water. In addition, there were significant correlations with the interaction between the national border and geographic distance, and the interaction between separation by water and geographic distance. The language-external factors accounted for approximately 45% of the variation.

The *Human body* Domain

Table 8 shows the MRM results for in the *Human body* domain. Main factors that were correlated with linguistic distance were geographic distance, dialect area, and the national border. There were no significant correlations with interactions between these factors. The language-external factors accounted for approximately 26% of the variation.

The *Society and Education* Domain

Table 9 shows the MRM results for in the *Society and education* domain. The only main factor that was correlated with linguistic distance was geographic distance. There were no significant correlations with interactions between the factors. The language-external factors accounted for approximately 9% of the variation.

Summary of the Four Domains

Table 10 summarizes the findings across the four semantic domains. The table lists which language-external factors significantly predicted linguistic distance in each domain (with "+" for positive coefficients and "–" for negative coefficients), indicates which of these was most strongly correlated based on the Partial Mantel correlations (shaded grey; see **Supplementary Materials**), and provides the R^2-values obtained through the MRM analyses. The table shows that there are both similarities and differences across domains.

Geographic distance significantly predicted linguistic distance in all four domains—Partial Mantel correlations ranged between r = 0.104 and r = 0.355; see **Supplementary Materials**. In fact, geographic distance was the strongest correlate with linguistic distance across all domains, which shows that linguistic differences within a relatively coherent dialect area such as the Limburgish one primarily arise from natural patterns of contact between communities.

Dialect area significantly predicted linguistic distance in three domains—not for the *Society and education* domain. As expected, linguistic distances were higher when locations are from different dialect areas, highlighting the role of smaller coherent subunits within an overall dialect area.

The national border significantly predicted linguistic distance in three of the four domains—again, not for the *Society and education* domain. The coefficients were always positive, confirming that the border acts as an additional barrier to contact between dialects. There was only one domain for which there was a significant interaction between the border and geographic distance, which indicates that the border generally acts as a barrier irrespective of distance. However, the fact that this interaction was negative seems to show that with increasing distance, the effect of the national border as a barrier can diminish. This was expected given that large distances between locations already hinder contact in themselves.

Separation by water significantly predicted linguistic distance in two of the four domains. Interestingly, the coefficients were always negative, indicating that dialects on separate sides of the Meuse river are more like each other. This finding might be counterintuitive at first, but in both cases, there was a positive significant interaction between separation by water and geographic distance, which we believe to be important in understanding this effect. The Meuse river provides an important means of transport in the area and upstream/downstream travel facilitates contact between towns on opposite sides even if they are relatively far apart. However, the further apart two locations are from each other, the more likely it is that neither of them is close to river at all, and the less likely it is that the Meuse facilitates contact between them, and so the positive interaction reverses its effect on linguistic distance.

In contrast to previous work that found population size to play an important role in linguistic variation (although see e.g., Nerbonne and

TABLE 10 | Significant explanatory factors (+ for positive coefficients; – for negative coeffecients) and R^2-values across the four semantic domains based on the multiple regression of distance matrices (MRM) analyses.

	Church and religion	Clothing and personal hygiene	Human body	Society and education
Log geographic distance	+	+	+	+
Dialect area	+	+	+	
National border	+	+	+	
Border * distance		–		
Separation by water	–	–		
Water * distance	+	+		
Log population difference				
MRM R^2	0.33	0.45	0.26	0.09

TABLE 11 | Linear mixed-effect modelling results for the *Church and religion* domain, showing beta coefficients, standard errors, t-values and significance levels.

	β	SE	*t*	*p*
Intercept	–0.015	0.045	0.34	0.731
Log geographic distance	0.299	0.009	34.39	<0.001
Dialect area	0.022	0.007	2.92	0.003
National border	0.352	0.008	42.99	<0.001
Border * distance	0.034	0.008	4.06	<0.001
Separation by water	0.097	0.007	13.72	<0.001
Water * distance	0.035	0.008	4.46	<0.001
Log population difference	0.012	0.009	1.38	0.167

Conditional R^2 = 0.538, Marginal R^2 = 0.314.

TABLE 12 | Linear mixed-effect modelling results for the *Clothing and personal hygiene* domain, showing beta coefficients, standard errors, t-values and significance levels.

	β	SE	*t*	*p*
Intercept	–0.036	0.027	1.33	0.183
Log geographic distance	0.460	0.005	86.59	<0.001
Dialect area	–0.014	0.004	3.22	0.001
National border	0.346	0.005	72.43	<0.001
Border * distance	0.136	0.005	25.60	<0.001
Separation by water	0.011	0.005	2.24	0.025
Water * distance	–0.020	0.005	3.99	<0.001
Log population difference	–0.003	0.006	0.46	0.645

Conditional R^2 = 0.546, Marginal R^2 = 0.420.

Heeringa, 2007), difference in population size between the two locations was not significantly correlated with linguistic distance in any of the domains investigated in this study.

Finally, results from the MRM analyses showed that language external factors accounted for between only 9% and up to 45% of the variance, showing that the predictive power of such external factors can differ considerably between domains.

Regression Analyses of Language-External Factors

In the second step of our analyses, we used linear mixed-effect modelling to further analyze the data. Critically, this approach allowed us to include location as a random variable in the analyses, thereby making it possible to account for differences in individual uniqueness of the locations included in our study. As with the previous analyses, we present the results on a domain basis, followed by a summary of the findings, and finally compare these results to what was found in the MRM analyses.

Language-External Factors in the *Church and Religion* Domain

Table 11 shows the results of the linear mixed-effect modelling for the *Church and religion* domain. Significant predictors of linguistic distance were geographic distance, dialect area, the national border, and separation by water. In addition, there were significant interaction effects between the national border and distance, as well as separation by water and distance. Overall, the model accounted for approximately 54% of the variance, of which 23% was accounted for by the inclusion of the random effect of location.

Language-External Factors in the *Clothing and Personal Hygiene* Domain

Table 12 shows the results of the linear mixed-effect modelling for the *Clothing and personal hygiene* domain. Significant predictors of linguistic distance were geographic distance, dialect area, the national border, and separation by water. In addition, there were significant interaction effects between the national border and distance, as well as between separation by water and distance. Overall, the model accounted for approximately 55% of the variance, of which 13% was accounted for by the inclusion of the random effect of location.

Language-External Factors in the *Human body* Domain

Table 13 shows the results of the linear mixed-effect modelling for the *Human body* domain. Significant predictors of linguistic distance were geographic distance, the national border, and separation by water. In addition, there were significant interaction effects between the national border and distance, as well as separation by water and distance. Overall, the model accounted for approximately 51% of the variance, of which 22% was accounted for by the inclusion of the random effect of location.

TABLE 13 | Linear mixed-effect modeling results for the *Human body* domain, showing beta coefficients, standard errors, t-values and significance levels.

	β	SE	*t*	*p*
Intercept	−0.015	0.037	0.41	0.681
Log geographic distance	0.383	0.006	64.31	<0.001
Dialect area	0.007	0.005	1.43	0.154
National border	0.241	0.005	47.00	<0.001
Border * distance	0.073	0.005	13.25	<0.001
Separation by water	0.097	0.005	2.68	<0.001
Water * distance	−0.012	0.005	2.42	0.016
Log population difference	0.011	0.006	1.86	0.063

Conditional R^2 = 0.510, Marginal R^2 = 0.286.

TABLE 14 | Linear mixed-effect modeling results for the *Society and education* domain, showing beta coefficients, standard errors, t-values and significance levels.

	β	SE	*t*	*p*
Intercept	−0.007	0.028	0.27	0.791
Log geographic distance	0.215	0.007	28.92	<0.001
Dialect area	0.008	0.006	1.33	0.183
National border	0.287	0.007	41.53	<0.001
Border * distance	0.024	0.007	3.24	0.001
Separation by water	0.007	0.006	1.08	0.280
Water * distance	0.004	0.007	0.64	0.522
Log population difference	0.028	0.008	3.67	<0.001

Conditional R^2 = 0.292, Marginal R^2 = 0.167.

Language-External Factors in the *Society and Education* Domain

Table 14 shows the results of the linear mixed-effect modelling for the *Society and education* domain. Significant predictors of linguistic distance were geographic distance, the national border, and population difference. In addition, there was a significant interaction effect between the national border and distance. Overall, the model accounted for approximately 29% of the variance, of which 12% was accounted for by the inclusion of the random effect of location.

Summary of Language-External Factors Across the Four Domains

Table 15 summarizes the findings across the four domains. The table lists which language-external factors were significant predictors of linguistic distance in each domain (with "+" for positive coefficients and "–" for negative coefficients), indicates which of these was the strongest predictors (shaded grey), and provides the conditional and marginal R^2-values obtained through the linear mixed-effect modelling.

A conspicuous and reassuring outcome is the similarity of the MRM R^2s in **Table 10** and the Marginal R^2s in **Table 15**. Both techniques approximately use the same sources of variation, but in the LMERs the random part (the random intercepts of the locations) is defined separately and excluded from the marginal R^2. The conditional R^2s are therefore higher, even much higher in our analyses, because of the relevance of individual dialect differences. At the same time, this has the consequence that we have more and stronger effects in the LMER analyses because they are related to the part of the variation defined as the conditional R^2s.

As found in the correlational analyses, geographic distance was a significant predictor of linguistic distance in all domains. In contrast to the correlational analyses however, geographic distance was the strongest predictor in only two of the four domains. Interestingly, these were the two least standardized domains (*Clothing and personal hygiene*, and *the Human body*), highlighting that patterns of linguistic variation develop naturally through contact when there is no additional homogenization that results from standardization processes.

Differences between dialect areas significantly predicted overall linguistic distances in only two of the four semantic domains. That the effect in one of these domains (*Clothing and personal hygiene*) was negative is puzzling, but its small effect size indicates that the effect is negligible ($\beta = -0.014$).

The national border was a significant predictor of linguistic distance in all semantic domains, and it was the strongest predictor in two domains. These were the more standardized domains (*Church and religion*, and *Society and education*), which shows that overall standardization differences between Belgium and the Netherlands are reflected in increased linguistic differences between Limburgish varieties from different countries. Contrary to what we originally expected, this was also the case for the *Church and religion* domain. In addition, all domains showed a significant positive interaction between the national border and geographic distance, indicating that these two factors work in tandem with increased linguistic distances as a result.

In contrast to the results reported above (*Summary of the Four Domains*), the mixed-effect regressions showed increasing linguistic distances between locations that were separated by water—which is more in line with what we initially expected. That this effect was significant in combination with the national border indicates that they are independent barriers to contact. In addition, there was a significant interaction effect between separation by water and geographic distance in the three domains for which there was a main effect. In two (less standardized) domains, this interaction was negative indicating that the effect of separation by water decreases over distance, but for the *Church and religion* domain, the two work in tandem to further increase linguistic distances between locations.

Differences in population size turned out significant in only one domain, which is largely in line with the lack of significant correlations found in *Multiple Regression Over Distance Matrices*. The positive value of this effect seems to indicate that there are large linguistic distances between communities of different sizes.

Finally, R^2-values of the mixed-effect models were on average around 20 percentage points higher than those for the MRM analyses. To assess the value of including location as a random variable, we compared the models presented above with models that did not include random effects, which showed that the

TABLE 15 | Significant predictors across the four semantic domains (strongest predictor highlighted), and conditional and marginal R^2-values for the linear mixed-effect models.

	Church and religion	Clothing and personal hygiene	Human body	Society and education
Log geographic distance	+	+	+	+
Dialect area	+	−		
National border	+	+	+	+
Border * distance	+	+	+	+
Separation by water	+	+	+	
Water * distance	+	−	−	
Log population difference				+
Conditional R^2	0.54	0.55	0.51	0.29
Marginal R^2	0.31	0.42	0.29	0.17

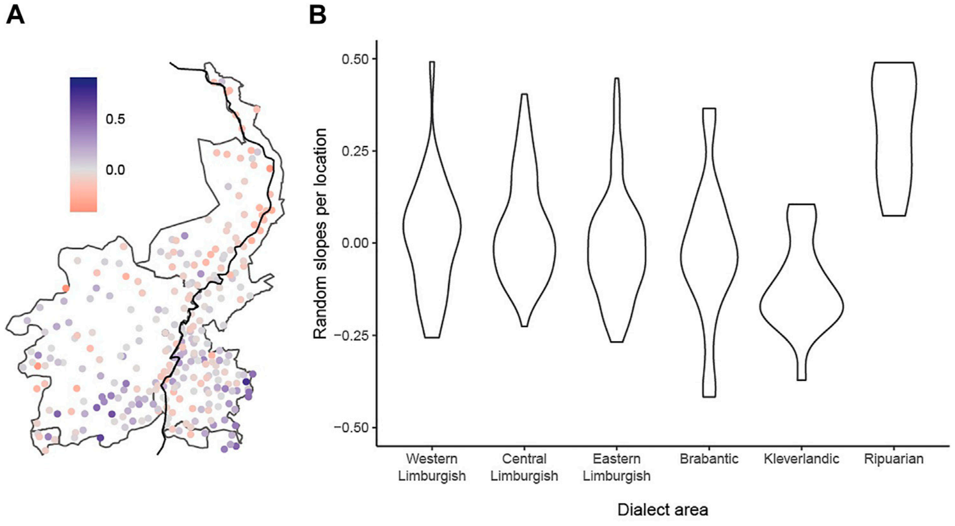


FIGURE 6 | Map showing the Limburgish area with values of the mean random intercept for each location **(A)**, and distribution of random intercepts across six dialect areas **(B)**.

random effect significantly improved the model for all domains (all p's < 0.001). Including location as a random effect accounted for between 12 and 23 percentage points across the domains.

Spatial Patterns in Dialect Uniqueness

Dialect uniqueness has been addressed before, e.g., by Jeszenszky et al. (2019: 17), who provide a map of Japonic varieties showing average linguistic distance toward all other survey locations. While their use of average distance was able to pick up the mixed nature of varieties spoken in Hokkaido, the average linguistic distances in their map seem to be highest in peripheral areas, which is to be expected given they're the furthest away from most other varieties. One way to better take this periphery component into consideration is to use the random intercepts of the LMER analysis. This makes it possible to further investigate spatial patterns in the random effect, i.e., spatial patterns in dialect uniqueness. To do so, we plotted all locations on a map of the Limburg area and colored them according to their mean random effect over the four semantic domains—see **Figure 6**, panel (a). In addition, we created a set of violin plots to show the distribution of the random intercepts across the six dialect areas—see **Figure 6**, panel (b).

As the figure shows, the random intercept for varieties in the three core areas (Western-, Central-, and Eastern Limburgish) center around zero and show a similar distribution in each area. The peripheral areas show more differences, however. While random intercepts for the locations in the Brabantic area show a pattern that is similar to the core Limburgish varieties, intercepts for the Kleverlandic varieties are skewed towards negative values (indicating smaller linguistic distances than expected), whereas intercepts for the Ripuarian varieties in the east, and some varieties in the southwest are skewed towards higher positives values (indicating larger linguistic differences than expected). For the Ripuarian data, these results are expected as the Ripuarian dialects are linguistically much closer to German dialects than the other dialects spoken in the Limburgish dialect area. Similarly, the varieties in the southwest have been shown to be influenced by French due to their proximity to the Germanic-Romance language border (van Hout, Kruijsen and Gerritsen, 2014), and this influence is particularly strong for the *Clothing and personal hygiene* domain (Franco et al., 2019a) included here. Finally, for the Kleverlandic dialects, the negative intercepts may perhaps be explained by the fact that these dialects are geographically the most outspoken edge of the Limburgish dialect area, where many very large geographical distances are expected to predict large linguistic distances. The negative intercepts seem to correct this peripheral overestimation as in e.g., Jeszenszky et al. (2019). Thus, the dialects that are spoken in the Kleverlandic region seem to resemble the language of the central regions more than expected on the basis of their location.

Including Language-Internal Factors

Our final step of the analyses comprised the inclusion of language-internal factors to the linear mixed-effect model. As described above (see above, *Language-Internal Factors*), we coded several characteristics for each semantic domain: 1) the number of subsections at different levels of depth, 2) the number of concepts, both in total and at different levels of depth, 3) the ratio of multi-word concepts, and 4) the mean and median length of the concept headword.

As many of the language-internal factors were highly correlated (see **Supplementary Table S10** for a complete overview), we conducted principal factor analysis to determine the structure of their common variance, which showed that our set of internal factors was computationally singular. As such, we merged all internal factors into a single variable based on the mean of their z-values.

We then conducted a series of linear mixed-effect regression analyses, which included all external factors as described in the previous section, with the addition of 1) semantic domain as a nominal variable (with *Society and education* as the reference level), 2) the single merged value for all internal factors combined, and 3) each language-internal factor individually. We compared these new models with the baseline model that only included language-external factors. A summary of these comparisons is shown in **Table 16**.

As the table shows, all individual language-internal factors significantly improved the model, indicating that there is added value in including language-internal factors when trying to model patterns of linguistic variation. For the individual factors, the two measures of concept headword length provided largest improvement—even more so than the merged value for all language-internal factors combined. Their positive betas confirm that for less salient concepts, larger linguistic distances are found across the Limburgish dialect area. At the same time however, the AIC values show that the addition of semantic domain as a nominal variable produces the best model, suggesting that there are additional domain-specific characteristics that were not captured by the language-internal factors here.

TABLE 16 | Overview of models including language-internal factors, showing beta coefficients, Akaike information criterion values compared to the baseline model with external factors only, χ^2-values, and significance levels.

	β	AIC	χ^2	p
External factors only		1,82,438		
Domain (nominal)				
Church and religion	0.118	1,81,087	1,383	<0.001
Clothing and personal hygiene	0.079			
Human body	0.036			
All internal factors merged	0.069	1,81,920	528.7	<0.001
Subsections at one level of depth	0.068	1,81,928	521.5	<0.001
Subsections at two levels depth	0.020	1,82,409	42.24	<0.001
Subsections at maximum depth	–0.024	1,82,388	67.13	<0.001
Total number of concepts	0.035	1,82,331	129.1	<0.001
Concepts at one level of depth	–0.006	1,82,452	4.31	0.038
Concepts at two levels depth	0.039	1,82,297	156.5	<0.001
Concepts at maximum depth	0.076	1,81,834	618.1	<0.001
Ratio of multi-word concepts	0.057	1,82,086	359.3	<0.001
Mean concept length	0.085	1,81,623	828.0	<0.001
Median concept length	0.078	1,81,764	686.5	<0.001

df for Domain as nominal variable = 3; all other df's = 1.

DISCUSSION

In this paper, we showcased spatial analysis techniques for dialect geography. After conducting a correlational analysis with Mantel correlations and MRM, we used linear mixed-effects regression (LMER) modelling to further investigate the effect of language-external factors while accounting for location-based variation by including it as a random factor. This method also allowed us to critically assess the importance of a set of independent variables that have been shown to affect processes of linguistics diffusion, both language-external and language-internal. All in all, our results confirm that geographic distance is a very important predictor of linguistic differences. However, depending on the semantic domain under scrutiny, other language-external factors were shown to play a significant role as well. For example, in the semantic domains of *Church and religion* and *Society and education*, our analyses revealed that the national border has a larger effect. Finally, our models improved when language-internal variables were included in the analysis, further confirming that linguistic distances differ between semantic domains, an observation that may be relevant for future work in lexical dialectology, as well as for lexical research more broadly.

There are a number of advantages of using the techniques proposed here over methods that have a longer standing in the field. First, comparing the results for the external factors obtained with MRM vis-à-vis our linear mixed-effects models shows that they are highly similar across the board, indicating that our LMER approach is a suitable technique for this line of research. Moreover, mixed-effects modelling makes it possible to incorporate the inherent uniqueness of individual locations by handling them as a random factor. While previous work has used individual locations as random effects (e.g., Wieling, 2012; Wieling et al., 2014; Wieling et al., 2018), these studies compared each location to only a single point of reference, and the random effect gives insight into how each dialect compares to the other dialects in its divergence from the standard. In contrast, the LMER approach described here uses linear models through which we can apply regression to *all* pairs of linguistic differences. As such, each location is compared with all other locations and the random effect provides the additional insight into the individual uniqueness of each dialect. In this study specifically, where we aimed to understand the relationship between linguistic and geographic distance, the use of random intercepts informs us of the position of individual locations (or groups of locations) in the overall linguistic area. While both correlational analyses also explore sources of variation, the random intercepts of locations in the LMERs are defined separately. The consequence of this is stronger and more clearly defined effects in the LMER analyses. Of course, another source of variation may be that, depending on the location, linguistic distances may increase at a higher or lower rate. Although we also examined the effect of such random slopes in the analyses, the models did not converge. We certainly need to further explore the many possibilities of linear mixed regression.

The LMER method proposed here shares with previous approaches (GAM; e.g., Wieling, 2012) the opportunity of incorporating data that cannot be coded into pairwise distance matrices, such as language-internal factors. This makes the method similar to other work in lectometry interested in the causes of variation in aggregate measures (Schneider, 1988; Pickl, 2013; Ruette and Speelman, 2013; Plevoets, 2020). Moreover, we might go even further and include the individual concepts as random factors (as in work using GAM), much as current approaches in psycholinguistic research (cf. Winter, 2019 for an introduction). In sum, the technique has large flexibility in handling random structures inherent to large data sets, allowing the researcher to systematically investigate latent sources of variation in their data. This is particularly relevant given the known importance of language-internal variables and general cognitive principles in linguistic variation (see the references in Franco et al., 2019b).

Some questions remain, however. First, our analyses used the Weighted Dissimilarity Value as a measure of linguistic distance, as its results were more regular than what we obtained for string edit distance. It is possible that the large number of data points, as well as long responses for a subset of concepts, may have resulted in large linguistic distances that contain unnecessary noise. More detailed investigation is needed to determine whether the use of string edit distance—as is common in dialectometry—becomes unstable in certain cases, e.g., for large datasets. Thus, follow-up research should investigate other measures for linguistic distance. One option worth exploring is weighing linguistic dissimilarity based on the geographic density of the lexical variants. Variants with a dense geographical distribution may prove more informative on the role of geography.

The current study uses straight line distances, but there are of course other ways of operationalizing geographic distance that have been used in studies on patterns of linguistic variation, such as travel distance (e.g., Inoue, 2004; Jeszenszky et al., 2019) and travel time (e.g., Gooskens, 2005; Jeszenszky et al., 2019), or using longitude/latitude (e.g., Wieling, 2012; Wieling et al., 2014; Wieling et al., 2018). For the Limburgish dialect area, we expect our results to stay the same when using such measures, as there are no other major geographical obstacles (e.g., mountain ranges, marshlands) that would further impede travel. For the whole of the Netherlands, straight line distance and travel distance have been shown to correlate strongly (r > 0.9; van Gemert, 2002). In fact, even in an area as mountainous as Japan, hiking distance and modern travel distance both correlate strongly with straight line distance (Jeszenszky et al., 2019). Nevertheless, Gooskens (2005) showed that incorporating *historical*—rather than modern—travel times produced better models for Norwegian varieties, so there are potential benefits for some linguistic areas provided that such historical data is available.

Another open question concerns the identification and importance of language-internal factors. While we were able to show that inclusion of language-internal factors improved our model, the largest improvement was obtained by simply including semantic domain as a nominal variable. It thus remains an open question which factors cause semantic domains to differ from each other. While we specifically chose four semantic domains that vary with regard to their degree of standardization and cultural variability, these interpretations do not unequivocally explain the results we obtained. For example, **Table 16** shows that, if language-external variables are controlled for, the β is the highest in the field C*hurch and religion*, which is the field where little variation would be expected. Further work looking at additional domains and subdomains is needed to better understand the role of different language-internal factors in the emergence and persistence of variation across a dialect area.

The approach taken in the current study might be qualified as *computational dialect geography*. In fact, we prefer this label over the one of *dialectometry*. Dialectometry perfectly fits the developments in the past, i.e., stipulating that dialect phenomena are measurable, but as new computational procedures and algorithms emerge and get applied, we believe there is a broader potential of handling and analyzing language variation data, including the many internal and external factors, giving the floor to computational dialect geography, i.e., computational sociolinguistics. These developments give room to additional next-level techniques such as machine learning and deep learning to research dialect classification problems, and computational intelligence to understand the trade-off between processes of convergence and divergence in short-term and long-term communicative processes. Future work can also build on methods that are used to optimize complex functions to better understand the functional relation between linguistic and geographic distance. Finally, the use of simulations in predicting linguistic variation is not new (see e.g., Hard, 1972), but incorporating techniques from other fields can move these attempts forward in testing and revealing the underlying parameters and processes of linguistic variation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/2jcxa/.

AUTHOR CONTRIBUTIONS

Conceptualization: JH, KF, RH, Data curation: KF, JH, Formal analysis: JH, Interpreting results and writing: JH, KF, RH. Revisions: JH, KF, RH.

FUNDING

This research was supported by Radboud University and KU Leuven.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.668035/full#supplementary-material

REFERENCES

Berlin, B., Breedlove, D. E., and Raven, P. H. (1973). General Principles of Classification and Nomenclature in Folk Biology. *Am. Anthropologist* 75 (1), 214–242. doi:10.1525/aa.1973.75.1.02a00140

Berlin, B. (1972). Speculations on the Growth of Ethnobotanical Nomenclature. *Lang. Soc.* 1 (1), 51–86. doi:10.1017/s0047404500006540

Berlin, B. (1978). “Ethnobiological Classification,” in *Cognition and Categorization*. Editors E. Rosch and B. B. Lloyd (New York: Wiley), 9–26.

Bloomfield, L. (19581933). *Language* (6th edition). London: George Allen & Unwin.

Cajot, J. (1977). De Rijksgrens Tussen Beide Limburgen Als Taalgrens. *Mededelingen van de Vereniging voor Limburgse Dialect- en Naamkunde* 4, 3–15.

CBS (2017). *Kerncijfers wijken en buurten 2017*. Den Haag: Centraal Bureau voor de Statistiek. Available at: https://www.cbs.nl/nl-nl/maatwerk/2017/31/kerncijfers-wijken-en-buurten-2017. (Data collected on March 26, 2019). (last accessed on 6 January, 2021).

CBS (2021). *Regionale Kerncijfers Nederland*. Den Haag: Centraal Bureau voor de Statistiek. Available at: https://opendata.cbs.nl/#/CBS/nl/dataset/70072ned/table?ts=1551669572059/. (Data collected on March 26, 2019). (last accessed on 6 January, 2021).

Chambers, J. K., and Trudgill, P. (1998). *Dialectology*. 2nd ed. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511805103

De Ceuninck, K. (2009). Lokale en regionale politiek: de gemeentelijke fusies van 1976: een mijlpaal voor de lokale besturen in België. Brugge: Vanden Broele.

Franco, K., Geeraerts, D., Speelman, D., and van Hout, R. (2019a). Maps, Meanings and Loanwords: The Interaction of Geography and Semantics in Lexical Borrowing. *J. Ling. Geogr.* 7 (1), 14–32. doi:10.1017/jlg.2019.2

Franco, K., Geeraerts, D., Speelman, D., and van Hout, R. (2019b). Concept Characteristics and Variation in Lexical Diversity in Two Dutch Dialect Areas. *Cogn. Linguistics* 30 (1), 205–242. doi:10.1515/cog-2017-0136

Geeraerts, D., and Speelman, D. (2010). “Heterodox Concept Features and Onomasiological Heterogeneity in Dialects,” in Advances In Cognitive Sociolinguistics *(Cognitive Linguistics Research 45)*. Editors D Geeraerts, G Kristiansen, and Y Peirsman (Berlin: De Gruyter Mouton), 23–39.

Geeraerts, D., Grondelaers, S., and Bakema, P. (1994). The Structure Of Lexical Variation: Meaning, Naming, and Context *(Cognitive Linguistics Research 5)*. Berlin: De Gruyter Mouton. doi:10.1515/9783110873061

Gerritsen, M., and van Hout, R. (2006). “Sociolinguistic Developments as a Diffusion Process,”. *Sociolinguistics: An International Handbook of the Science of Language and Society/Soziolinguistik: Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*. Editors U. Ammon, N. Dittmar, K. J. Mattheier, and P. Trudgill. 2nd ed. (Berlin: De Gruyter Mouton), Vol. 3.

Gerritsen, M. (1999). Divergence of Dialects in a Linguistic Laboratory Near the Belgian-Dutch-German Border: Similar Dialects under the Influence of Different Standard Languages. *Lang. Change* 11 (1), 43–65. doi:10.1017/s0954394599111037

Goebl, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Max Niemeyer.

Goebl, H. (2006). Recent Advances in Salzburg Dialectometry. *Literary linguistic Comput.* 21 (4), 411–435. doi:10.1093/llc/fql042

Gooskens, C. (2005). Traveling Time as a Predictor of Linguistic Distance. *Dialectologia et Geolinguistica* 13, 38–62.

Goslee, S. C., and Urban, D. L. (2007). The Ecodist Package for Dissimilarity-Based Analysis of Ecological Data. *J. Stat. Softw.* 22 (7), 1–19. doi:10.18637/jss.v022.i07

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A Statistical Method for the Identification and Aggregation of Regional Linguistic Variation. *Lang. Change* 23 (2), 193–221. doi:10.1017/s095439451100007x

Hard, G. (1972). “Ein geographisches Simulationsmodell für die rheinische Sprachgeschichte,” in *Festschrift Matthias Zender, Studien Zur Volkskultur*. Editors E. Essen and G. Wiegelmann (Bonn: Sprache und Landesgeschichte), 5–29.

Heeringa, W., and Nerbonne, J. (2001). Dialect Areas and Dialect Continua. *Lang. Change* 13 (3), 375–400. doi:10.1017/s0954394501133041

Hinskens, F., Kallen, J. L., and Taeldeman, J. (2000). Merging and Drifting Apart. Convergence and Divergence of Dialects across Political Borders. *Int. J. Sociol. Lang.* 145, 1–28. doi:10.1515/ijsl.2000.145.1

Huisman, J. L. A., Majid, A., and van Hout, R. (2019). The Geographical Configuration of a Language Area Influences Linguistic diversity. *PLoS ONE* 14 (6), e0217363. doi:10.1371/journal.pone.0217363

Inoue, F. (2004). Geographical Factors of Communication on the Basis of Usage Rate of the Standard Japanese Forms and Railway Distance. *Jpn. J. Lang. Soc.* 7 (1), 19–29. doi:10.19024/jajls.7.1_19

Jeszenszky, P., Hikosaka, Y., Imamura, S., and Yano, K. (2019). Japanese Lexical Variation Explained by Spatial Contact Patterns. *Ijgi* 8 (9), 400. doi:10.3390/ijgi8090400

Kemp, C., Xu, Y., and Regier, T. (2018). Semantic Typology and Efficient Communication. *Annu. Rev. Linguist.* 4, 109–128. doi:10.1146/annurev-linguistics-011817-045406

Ko, V., Wieling, M., Wit, E., Nerbonne, J., and Krijnen, W. (2014). Social, Geographical, and Lexical Influences on Dutch Dialect Pronunciations. *Comput. Linguistics Neth. J.* 4, 29–38.

Labov, W. (2007). Transmission and Diffusion. *Language* 83 (2), 344–387. doi:10.1353/lan.2007.0082

Langacker, R. (1987). Foundations of Cognitive Grammar,” in *Theoretical Prerequisites* (Stanford: Stanford University Press), Vol. 1.

Lee, S., and Hasegawa, T. (2014). Oceanic Barriers Promote Language Diversification in the Japanese Islands. *J. Evol. Biol.* 27 (9), 1905–1912. doi:10.1111/jeb.12442

Legendre, P., and Legendre, L. (2012). *Numerical Ecology*. New York: Elsevier.

Lichstein, J. W. (2007). Multiple Regression on Distance Matrices: a Multivariate Spatial Analysis Tool. *Plant Ecol.* 188 (2), 117–131. doi:10.1007/s11258-006-9126-3

Majid, A., and Burenhult, N. (2014). Odors Are Expressible in Language, as Long as You Speak the Right Language. *Cognition* 130 (2), 266–270. doi:10.1016/j.cognition.2013.11.004

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Res.* 27 (2), 209–220.

Nerbonne, J., and Heeringa, W. (2001). Dialect Areas and Dialect Continua. *Lang. Variation Change* 13, 375–400.

Nerbonne, J., and Heeringa, W. (2007). “Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation,” in *Roots: Linguistics in Search of* *its Evidential Base*. Editors S. Featherston and W. Sternefeld (Berlin: Mouton de Gruyter), 267–297.

Nerbonne, J., and Kleiweg, P. (2007). Toward a Dialectological Yardstick*. *J. Quantitative Linguistics* 14 (2–3), 148–166. doi:10.1080/09296170701379260

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T. (2011). Gabmap - A Web Application for Dialectology. *Dialectologia Special Issue II*, 65–89.

Nerbonne, J. (2010). Measuring the Diffusion of Linguistic Change. *Phil. Trans. R. Soc. B* 365 (1559), 3821–3828. doi:10.1098/rstb.2010.0048

Nerbonne, J. (2013). "How Much Does Geography Influence Language Variation?," in *Space in Language and Linguistics. Geographical, Interactional, and Cognitive Perspectives*. Editors, et al. (Berlin:De Gruyter), 220–236.

Pickl, S. (2013). Lexical Meaning and Spatial Distribution. Evidence from Geostatistical Dialectometry. *Literary Linguistic Comput.* 28 (1), 63–81. doi:10.1093/llc/fqs050

Plevoets, K. (2020). Lectometry and Latent Variables: a Model for Underlying Determinants of (Normative) Choices in Written and Audiovisual Translations. in: *Extending The Scope Of Lectometr*. Editors L. Rosseel, K. Franco, and M. Röthlisberger. *Special Issue of* Zeitschrift, 87 (2), 144–172. doi:10.25162/zdl-2020-0006

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic Objects in Natural Categories. *Cogn. Psychol.* 8 (3), 382–439. doi:10.1016/0010-0285(76)90013-x

Ruette, T., and Speelman, D. (2013). Transparent Aggregation of Variables with Individual Differences Scaling. *Literary Linguistic Comput.* 29 (1), 89–106. doi:10.1093/llc/fqt011

Schmeets, H. (2014). *De Religieuze Kaart Van Nederland, 2010-2013*. Centraal Bureau voor de Statistiek. Available at: https://www.cbs.nl/nl-nl/achtergrond/2014/40/de-religieuze-kaart-van-nederland-2010-2013 (Last accessed on 01 4, 2021).

Schneider, E. (1988). Qualitative vs. Quantitative Methods of Area Delimitation in Dialectology: a Comparison Based on Lexical Data from Georgia and Alabama. *J. English Linguistics* 21, 175–212.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35 (138), 335–357.

Shackleton, R. G. (2005). English-American Speech Relationships. *J. English Linguistics* 33, 99–160. doi:10.1177/0075424205279017

Speelman, D., and Geeraerts, D. (2008). The Role of Concept Characteristics in Lexical Dialectometry. *Int. J. Humanities Arts Comput.* 2 (1-2), 221–242. doi:10.3366/e1753854809000408

Stat Bel (2021). *Bevolking naar woonplaats, nationaliteit (Belg/niet-Belg), burgerlijke staat, leeftijd en geslacht.* Algemene Directie Statistiek - Statistics Belgium. Available at: https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?view=1b9e219b-0387-4a70-880a-dc5eccaa244c. (Data collected on March 26, 2019). (last accessed on 6 January, 2021)

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistic Dating. *Int. J. Am. Linguistics* 21 (2), 121–137. doi:10.1086/464321

Tadmor, U. (2009). "Loanwords in the World's Languages: Findings and Results," in *Martin Haspelmath & Uri Tadmor, Loanwords In the World's Languages*. Berlin & Boston: De Gruyter Mouton. 55–75. doi:10.1515/9783110218442.55

Trudgill, P. (1974). Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Lang. Soc.* 3 (2), 215–246. doi:10.1017/s0047404500004358

Van de Wijngaard, T., and Keulen, R. (2007). "De indeling van de Limburgse dialecten," in Riek van Klank: Inleiding in de Limburgse Dialecten *(Veldeke Taalstudies 2)*. Editors R. Keulen, T. Van de Wijngaard, H. Crompvoets, and F. Walraven (Sittard: Veldeke Limburg), 15–23.

van Gemert, I. (2002). *Het Geografisch Verklaren Van Dialectafstanden Met Een Geografisch Informatiesysteem (GIS)*. Rijksuniversiteit Groningen: Master's thesis

van Hout, R., Kruijsen, J., and Gerritsen, M. (2014). "Exosmosis along the Romance-Germanic Language Border in Belgium. The Diffusion of French Borrowings in the Dutch Dialects of Haspengouw," in *Ens queda la paraula: Estudis de lingüística aplicada en honor a M. Teresa Turell, 197-223*. Editors R. Casesnoves-Ferrer, M. Forcadell Guinjoan, and N. Gavaldà-Ferré (Barcelona: Institut universitari de lingüistica aplicada Universitat Pompeu Fabra).

Weijnen, A. A., Goossens, J., and Goossens, P. (1983). "Inleiding," in *Woordenboek van de Limburgse Dialecten: Inleiding & I. Agrarische Terminologie, Aflevering 1: Akker- en Weidegrond*. Editors A A. Weijnen, Jan. Goossens, and P. Goossens (Assen: Van Gorcum), 1–77.

Weijnen, A. A. (1966). Nederlandse Dialectkunde *(Studia Theodisca 10)*. Assen: Van Gorcum.

Wieling, M., and Nerbonne, J. (2015). Advances in Dialectometry. *Annu. Rev. Linguist.* 1 (1), 243–264. doi:10.1146/annurev-linguist-030514-124930

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical Differences between Tuscan Dialects and Standard Italian: Accounting for Geographic and Sociodemographic Variation Using Generalized Additive Mixed Modeling. *Language* 90 (3), 669–692. doi:10.1353/lan.2014.0064

Wieling, M., Valls, E., Baayen, R. H., and Nerbonne, J. (2018). "Border Effects Among Catalan Dialects," in *Mixed Effects Regression Models in Linguistics Springer: Quantitative Methods in the Humanities and Social Sciences*, 71–97. doi:10.1007/978-3-319-69830-4_5

Wieling, M. (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*. Groningen: Diss., University of Groningen.

Willemyns, R. (2013). *Dutch: Biography of a Language*. Oxford: Oxford University Press.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. New York: Routledge. doi:10.4324/9781315165547

Zuur, A., Ieno, E. N., and Smith, G. M. (2007). *Analyzing Ecological Data*. New York: Springer. doi:10.1007/978-0-387-45972-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/frai.2021.643770

Characterising Online News Comments: A Multi-Dimensional Cruise Through Online Registers

Katharina Ehret and Maite Taboada**

Discourse Processing Lab, Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada

News organisations often allow public comments at the bottom of their news stories. These comments constitute a fruitful source of data to investigate linguistic variation online; their characteristics, however, are rather understudied. This paper thus contributes to the description of online news comments and online language in English. In this spirit, we apply multi-dimensional analysis to a large dataset of online news comments and compare them to a corpus of online registers, thus placing online comments in the space of register variation online. We find that online news comments are involved-evaluative and informational at the same time, but mostly argumentative in nature, with such argumentation taking an informal shape. Our analyses lead us to conclude that online registers are a different mode of communication, neither spoken nor written, with individual variation across different types of online registers.

Keywords: register variation, online news comments, online language, multi-dimensional analysis, corpus linguistics

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Samira Shaikh,
University of North Carolina at Charlotte, United States
Heike Wiese,
Humboldt University of Berlin, Germany
Isobelle Clarke,
Lancaster University, United Kingdom

****Correspondence:***
Katharina Ehret
kehret@sfu.ca
Maite Taboada
mtaboada@sfu.ca

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *18 December 2020*
Accepted: *21 May 2021*
Published: *14 June 2021*

Citation:
Ehret K and Taboada M (2021) Characterising Online News Comments: A Multi-Dimensional Cruise Through Online Registers. Front. Artif. Intell. 4:643770. doi: 10.3389/frai.2021.643770

1 INTRODUCTION

We present a text-linguistic study of the characteristics of online news comments as compared to other online registers. In contrast to many other registers on the web, online news comments have so far not been thoroughly scrutinised. However, there has been a sense, among journalists (Woollaston, 2013; McGuire, 2015) and researchers alike (Godes and Mayzlin, 2004; Marcoccia, 2004; North, 2007), that online news comments are like conversation or dialogue. We have challenged this assumption, in a related article comparing online news comments to face-to-face conversation and other traditional registers: While online news comments were found to contain features of personal involvement typical of face-to-face conversation, they can best be described as a type of written, evaluative discourse (Ehret and Taboada, 2020). As a matter of fact, we argue that online news comments should be regarded as their own register, and that language on the web, in general, is quite different from either standard written or spoken language (Ehret and Taboada, 2020, 23–24). It is natural to describe new registers in terms of other, more familiar registers, which is perhaps what leads to the characterisation of online news comments as conversations. This label has also sometimes been applied to blogs, but has also been found inadequate, as Peterson (2011) has argued. In his analysis of blogs, Peterson found that, although blogs have an expressive potential, such potential is not realised in the same way as it is in conversation.

An ever-growing body of research analyses online language in general (e.g., Crystal, 2011; McCulloch, 2020), specific online registers, such as email (Frehner, 2008; McVeigh, 2020), blogs (Herring et al., 2004; Peterson, 2011), reviews (Taboada, 2011; Vásquez, 2014), Facebook (West, 2013; Farina, 2018), Twitter (Zappavigna, 2012; Clarke and Grieve, 2019), or online and social media language in general (Giltrow and Stein, 2009; Titak and Roberson, 2013; Page et al., 2014; Biber and

Egbert, 2016; Berber Sardinha, 2018; Biber and Egbert, 2018). Little attention, however, has been paid to the linguistic characteristics of online news comments, a register now ubiquitous in our interactions with news, whether on the pages of newspapers or through platforms such as Twitter and Facebook.

Against this backdrop, the present paper explores the structural linguistic properties of online news comments in comparison with other online registers such as travel and opinion blogs, interactive discussions and news reports, or advice pieces, since our previous analysis involved a traditional written and spoken corpus. We will thus establish what—if not like spontaneous conversation—online news comments are like in the context of other online registers. The data for our analysis is drawn from the comments section of the *Simon Fraser University Opinion and Comments Corpus* (SOCC) on the one hand, and the *Corpus of Online Registers of English* (CORE) on the other. SOCC is the largest corpus of online comments publicly available, while CORE is to date the largest available corpus of registers on the web. Methodologically, we conduct a multi-dimensional analysis (Biber, 1988), considering a comprehensive set of well-established lexico-grammatical features, to describe online news comments along the emerging dimensions of variation in our dataset.

Our analysis shows that multi-dimensional analysis (MDA) is very well suited to capturing the variation found in some common online registers. By applying the part-of-speech tag frequency statistics and dimensionality reduction characteristic of MDA, we are able to place online news comments in a unique space as compared to other online registers. To be more precise, we find that there are three dimensions along which online news comments can be described in online variational space, with two of them being most prominent. The first dimension, which we labelled "Involved-evaluative" points to the involved nature (in the Biberian sense; **Section 3**) of online registers and online comments, with an involvement that includes evaluative meaning. We find, however, that the most characteristic dimension is "Informational-argumentative", marked by information density (nominalisations, longer words) and argumentative features such as conjuncts. Finally, the third, minor dimension, "Narrative-descriptive vs. instructional" supports our analyses of the first two, showing an involved personal narrative mixed with instructional detail.

The paper is structured as follows: **Section 2** describes the data source and methodology. In **Section 3** the results of the MDA analysis are presented. **Section 4** discusses online news comments in light of the results. **Section 5** offers a brief summary and concluding remarks.

2 MATERIAL AND METHODS

2.1 Online News Comments and Other Online Registers

Our aim is to compare online comments to other, well-studied online registers. To that end, we use the *Corpus of Online Registers of English* (CORE), the largest, most diverse corpus of online language currently available (Biber et al., 2015; Egbert et al., 2015; Biber and Egbert, 2018). CORE was conceived as an attempt to classify various online registers. The data was obtained by sampling publicly-available documents and tagging them in a bottom-up process. About 50,000 web documents were labelled through crowd sourcing, resulting in six general (written) register types and several sub-registers. The general registers were provided by the researchers, but the sub-registers were crowd sourced and labelled by users according to guidelines (Biber et al., 2015). Registers were labelled according to their communicative purpose: to narrate events, describe or explain information, express opinion, persuade, explain instructions, or to express oneself through lyrics. Many of the sub-registers were deemed to be hybrid, because they include characteristics of more than one register or sub-register. CORE thus comprises, for instance, sub-registers (with main register in parentheses) such as personal blog (narrative), FAQ (description), review (opinion), description for sale[1] (persuasive), recipe (instructional), or song lyrics (lyrical).

TABLE 1 | Overview of analysed registers, corpus source, and number of words.

Register	Sub-register	Corpus	Word count
Narrative	Personal blog	CORE	3,264,463
—	Travel blog	CORE	382,124
—	Sports report	CORE	2,729,925
—	News report/blog	CORE	9,806,239
Informational description	FAQ	CORE	678,562
—	Description of a person	CORE	958,925
—	Informational blog	CORE	2,141,271
—	Encyclopedia article	CORE	1,613,338
—	Research article	CORE	1,905,846
Opinion	Opinion blog	CORE	10,898,872
—	Advice	CORE	1,415,912
—	Religious sermon/blog	CORE	1,435,058
—	Review	CORE	2,121,213
Persuasive	Description for sale	CORE	1,130,813
Instructional	Recipe	CORE	89,513
Interactive discussion	Interactive discussion	CORE	3,099,725
Online news comments	—	SOCC	5,779,157
Total	—	—	**49,450,956**

We chose CORE because of its focus on the public web, the readily available set of registers that one is likely to encounter online. An additional set of computer-mediated communication exists, including text messages (SMS, WhatsApp, Telegram, Signal, Direct Messages on Facebook or Twitter, etc.,), but that tends to be a one-to-one or small-group type of communication, not one to be publicly displayed the way online news comments are.

From this varied source of online materials, we select a large sample, excluding registers that are not unambiguously defined or not directly comparable to the online news comments we are interested in. In this vein, we exclude all hybrid registers, registers labelled as "other", lyrical and fully narrative registers, i.e. short story, historical article, and narrative, as well as spoken material. The sample does include typical online registers such as personal blog, travel blog, or news report,

[1] 'Description for sale' was originally labelled in CORE as 'Description with intent to sell'. We have shortened the label.

which are also labelled as narrative in CORE. In general, the sampling criterion excluded registers that may appear outside of the internet (short stories), but included online-only registers (travel blog), even when they were both under the same macro-register (narrative). This sample of CORE amounts to 43.7 million words (**Table 1**).

The online news comments come from the comments section of the *SFU Opinion and Comments Corpus* (SOCC), a large dataset of comments posted on the website of the Canadian English-language newspaper *The Globe and Mail* (Kolhatkar et al., 2020). The corpus contains more than 660,000 comments, a rough total of 37 million words. In this paper we specifically analyse comment threads, sequentially posted comments with a seemingly conversation-like structure, rather than individual comments. The analysis is furthermore restricted to comment threads with a minimum of 700 words, to improve the robustness of the multi-dimensional analysis (cf. Ehret and Taboada, 2020, 6). The comment threads were then analysed as individual comments, for a total of 5,949 comments. This selection of the SOCC corpus contains 5.8 million words and 388,141 sentences (but note that sentence boundaries are imprecise due to the online and informal nature of the data).

We should point out that we analyse comment threads rather than individual comments. This is in part due to technical considerations, because multi-dimensional analysis requires texts of a certain length, with 400 words the most common minimum length in the literature (Biber, 1995). There are also methodological considerations, in that what we are studying is the nature of online comments, which are typically posted in sequential form and constitute a thread of ideas and contributions. The drawback of this method is that the communicative function of one comment may be different from the next comment. We treat the entire thread as a communicative event, just like spoken conversations which include more than one participant.

2.2 Multi-Dimensional Analysis

Multi-dimensional analysis (MDA), originally introduced by Biber (1988) to describe variation in written and spoken registers of English, is a multi-variate statistical technique and the classic tool in text-linguistic approaches to register variation. MDA employs exploratory factor analysis to determine the shared variation in a given dataset based on the co-occurrence frequencies of linguistic features. The extracted factors are then interpreted as dimensions of variation according to the functional-communicative properties of the most important linguistic features on each factor.

We conduct a multi-dimensional analysis of our dataset largely following the statistical recommendations outlined in Biber (1988, 71–93), which we have also employed and detailed in previous work (Ehret and Taboada, 2020, 7–11). This paper differs from our previous work in that it focuses specifically on online language. To be more precise about the methodology, we apply maximum likelihood factor analysis as available in the R *stats* package and utilise a promax factor rotation. All statistics, unless otherwise indicated, were performed in R (R Core Team, 2020). The scripts, all statistics described here and elsewhere in the paper, additional statistical material, and data are available on GitHub.[2]

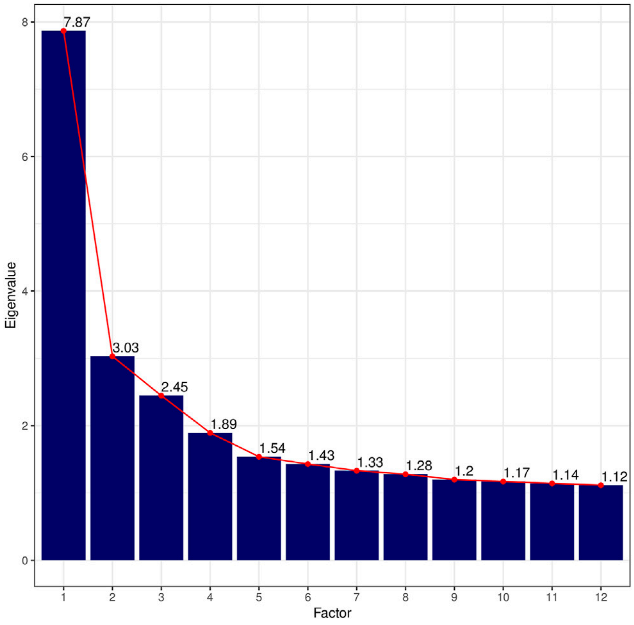


FIGURE 1 | Screeplot of eigenvalues for the first twelve factors. Eigenvalues were rounded to the second digit.

The linguistic features analysed in this paper consist of 67 core grammatical features of English customarily utilised in MDA studies (Biber, 1988; Biber and Finegan, 1989; Pavalanathan et al., 2017; Clarke and Grieve, 2019). These features include, but are not limited to, modals, pronouns, subordination and coordination, tense and aspect markers, as well as some special verb classes (Biber, 1988, 221–245). The dataset was automatically annotated with part-of-speech tags for these features using the *Multi-dimensional Analysis Tagger*, version 1.3.2 (Nini, 2019), a replication of Biber's (1988) original MDA tagger.[3] The part-of-speech tags and corresponding features are listed in **Supplementary Table S1** in the supplementary material. Subsequently, the occurrence frequencies of the 67 features were automatically retrieved, and normalised per 1,000 word tokens using a custom-made python script (available from our GitHub repository; see Data Availability Statement at the end of the paper). The features type-token-ratio (TTR) and average word length (AWL) were not normalised. Type-token ratio was calculated for the first 400 words in each text file, and average word length was calculated by dividing the number of orthographic characters by the number of tokens in each text file.

With an overall measure for sample adequacy of 0.77 and a *p*-value = 0 for *Bartlett's Test of Sphericity*, our dataset is statistically suitable for conducting a factor analysis (Dziuban

[2] https://github.com/sfu-discourse-lab/MDA-OnlineRegisters

[3] The tagger is based on and requires the Stanford part-of-speech tagger (Toutanova et al., 2003).

TABLE 2 | Mean factor scores per register. Positive values indicate that a register weighs on the positive pole of a factor; negative values indicate that a register weighs on the negative pole of a factor. All values were rounded to four decimal points.

Register	Factor 1	Factor 2	Factor 3	Factor 4
Advice	0.6092	−0.9682	0.3232	0.1130
Comments	0.1365	−0.0955	0.4659	0.4057
Description of a person	−0.5195	0.8685	−0.4012	−0.7035
Interactive discussion	1.1239	−0.2665	−0.3428	−0.0081
Encyclopedia article	−0.9321	0.2353	0.1470	−0.5167
FAQ	−0.4068	−1.1419	0.6247	0.1849
Informational blog	−0.4021	−0.6828	0.3098	0.2375
Description for sale	−0.4454	−0.8685	−0.3528	0.3339
News report	−0.5372	0.3953	−0.0345	−0.1031
Opinion blog	0.0520	−0.1194	0.2703	0.0676
Personal blog	0.9369	0.2646	−0.3415	−0.5101
Recipe	0.5258	-0.8007	−1.1673	−0.2321
Religious sermon	0.2640	0.1920	0.1669	−0.1248
Research article	−1.7230	−0.0744	1.2602	−0.1149
Review	0.1603	−0.3214	−0.4674	0.3110
Sports report	0.2669	0.3684	−0.8110	−0.2196
Travel blog	0.2887	0.0542	−0.6574	−0.5771

and Shirkey, 1974, 358–359). After inspecting the screeplot of eigenvalues in **Figure 1**, which shows a first break after the third factor before flattening out into a straight line, and the linguistic interpretability of the factors, we extract three factors for the final model (**Supplementary Table S2**). Traditionally, a factor is regarded as linguistically interpretable if it comprises at least five salient loadings. Following Biber (1988, 87), we consider loadings with a conservative cut-off ≥ |0.3| as statistically significant and hence salient. Note that Factor 3 is not fully linguistically interpretable according to these criteria, because it only comprises four salient loadings. However, it is included in the final model in order to avoid conflating factors, and to enhance the interpretability of the other factors in the model. Furthermore, for a tentative interpretation of Factor 3, we consider secondary features with loadings ≥ |0.2|. The total variance explained by the final model is about 20%.[4]

Finally, factor scores are automatically calculated for each text in the dataset. Factor scores indicate the position of each text on a given factor: the higher the absolute value of a factor score for a given text on a specific factor, the more typical is this text for the factor and the underlying linguistic dimension represented by the factor (Biber, 1988, 93). Additionally, factor scores also indicate on which pole of a factor a given text is to be positioned. Positive factor scores indicate that a given text weighs on the positive pole of a specific factor while negative scores indicate that a given text weighs on the negative pole of a specific factor. Consider, for instance, the text with the filename 19_N_personal_1747770_MAT.txt which belongs to the register personal blog. This text has a factor score of 2.36 on the first factor and a factor score of −0.85 on the second factor. On the basis of these factor scores, we can conclude that this text is more typical of Factor 1 than of Factor 2. Furthermore, the text contains many of the linguistic features which load high on the positive pole of Factor 1 and is marked by the absence of linguistic features which load high on Factor 2 (a detailed interpretation of the factors is given in **Section 3**).

In addition to factor scores, we calculate scores to position the individual registers as a whole on each factor. These scores are referred to as "mean factor scores" in this paper and are calculated as the arithmetic mean of the factor scores for all texts pertaining to a given register (**Table 2**).

3 DIMENSIONS OF LINGUISTIC VARIATION ONLINE

In this section, the extracted factors are interpreted as dimensions of variation. This means that each factor is linguistically interpreted based on the co-occurrence and complementary distribution of linguistic features and their shared functional-communicative properties (Biber, 1988, 91–92). Specifically, features with loading |≥ 0.3| are given the greatest importance in this interpretation, yet secondary features with less salient loadings are also considered. Features which load on multiple factors with the same polarity are primarily considered on the factor where they load highest. This interpretation is aided and confirmed by analysing the distribution of registers across the various dimensions. **Table 3** provides a summary of the three factors (for a complete list of features and loadings, see the GitHub repository in footnote 2).

The factors in our analysis and the emerging dimensions for this particular set of online registers vary from those that have been proposed for the CORE corpus by Biber and Egbert (2018). In their analysis, Biber and Egbert explore the entire CORE corpus, which, as we mention in **Section 2.1**, includes hybrid registers and spoken registers. Their first dimension, for instance, is thus "Oral-involved vs. literate", which captures the differences between song lyrics, TV dialogue, and interactive discussions on the one hand, and written registers such as research articles and encyclopedia entries on the other. Our dataset is a different one and, consequently, the emerging dimensions capture variation of online registers that are closer in nature to online news comments.

Factor 1 comprises 15 positive and seven negative features with salient loadings ≥ |0.3| and is therefore the most clearly defined factor. On the positive pole of the factor, we find features which are typical of spontaneous, informal, and involved communication such as contractions, first and second person pronouns, analytic negation, the pronoun *it*, private verbs which express personal attitudes or emotions (e.g. *believe*, *decide*, *know*), and emphatics (Biber, 1988, 105–106). In addition, some of the most salient features are not only well known as characteristic of spontaneous spoken language (Biber, 1988, 228–229), but have also been recently identified as markers of evaluation and opinion in online news comments (Ehret and Taboada, 2020, 13): *be* as main verb, adverbs, and predicative adjectives. Together, these three features often occur in constructions which are typically

[4] This would be considerably low if our primary interest was in variable reduction. However, the focus here is on the interpretability of the factors and the description of online news comments.

TABLE 3 | Overview of the three factors including features with loadings ≥ |0.3|. Positive loadings indicate co-occurrence of the features; negative loadings indicate complementary distribution.

Factor 1 **Involved-evaluative**		**Factor 2** **Informational-argumentative**	
Contractions	0.735	Nominalisations	0.716
First person pronouns	0.708	Average word length	0.652
Adverbs	0.599	THAT verb complement	0.355
Analytic negation	0.571	Conjuncts	0.347
Present tense	0.555	Attributive adjectives	0.319
BE as main verb	0.547	—	—
Pronoun IT	0.484	No negative features	—
Private verbs	0.46		
Emphatics	0.449		
Second person pronouns	0.445	**Factor 3**	
Conditional subordinator	0.423	**Narrative-descriptive vs. instructional**	
DO as proverb	0.398	Past tense	0.983
Predicative adjectives	0.35	Third person pronouns	0.375
THAT deletion	0.334	Public verbs	0.321
Demonstrative pronouns	0.33	—	—
—	—	Present tense	−0.523
Average word length	−1.036		
Nouns	−0.737		
Nominalisations	−0.706		
Prepositions	−0.64		
Attributive adjectives	−0.497		
Phrasal coordination	−0.462		
Past participle WHIZ deletion	−0.379		

used to convey evaluation (White, 2003; Hunston, 2011), such as in Example (1).

(1) a. It's$_{\text{be main verb}}$ not ideal$_{\text{predicative adjective}}$ for my husband [. . .] (personal blog, 19_N_personal _0000263_MAT.txt).
b. This is$_{\text{be main verb}}$ sometimes$_{\text{adverb}}$ hard$_{\text{predicative adjective}}$ to conjure up when you have been woken numerous times in the night to feed. (advice, 10_O_advice_3360949_MAT.txt).

On the negative pole of Factor 1, we find features which are well known as characteristic of an informational and abstract style in English: average word length, nouns, nominalisations, attributive adjectives, and prepositions are all indicators of information density and lexical specificity and are common in scientific or academic writing (Biber, 1988, 104–105). All in all, Factor 1 strongly resembles the Dimension "Involved vs. informational production" identified in Biber (1988) with, one could argue, an evaluative slant. We therefore interpret Factor 1 as Dimension 1 "Involved-evaluative vs. informational" and we shorten it to "Involved-evaluative" in the rest of the paper. In work by Biber and colleagues, multiple registers across different languages have been shown to be distributed across two main axes, involved vs. informational. The involved dimension refers to language use that includes "affective, interactional, and generalized content", as opposed to language with "high informational density and exact informational content" (Biber, 1988, p. 107).

This interpretation for Factor 1 dovetails with the distribution of registers on Dimension 1 (**Figure 2**). For instance, research and encyclopedia articles are located on the negative pole, while personal blogs and interactive discussions are representative of the positive pole of Dimension 1. Note that, in contrast to Biber's original Dimension 1, the dimension presented in this paper does not represent the fundamental distinction between written and spoken language. Instead, all registers analysed in this paper are written, and Dimension 1 thus distinguishes between online written discourse which is involved and evaluative and online written discourse which is informational (and presumably constructed as objective).

In contrast to the first factor, Factor 2 is defined exclusively by positive features. The five salient positive features are nominalisations, average word length, *that* verb complement, conjuncts, and attributive adjectives. The co-occurrence of nominalisations, high average word length, conjuncts, and attributive adjectives are indicators of information density and information integration. Nominalisations can also be interpreted as conveying specialised or abstract information (Biber, 1988, 227) such as, for instance, in scientific discourse. Conjuncts (e.g. *however*, *on the other hand*) are also prominent markers of argumentation and coherence (Halliday and Hasan, 1976; van Eemeren et al., 2007; Tseronis, 2011; Kolhatkar and Taboada, 2017b) as exemplified in (2-a). The argumentative aspect of Factor 2 is stressed by the secondary non-salient feature suasive verbs (feature loading 0.296) which express varying degrees of persuasion such as *propose*, *suggest*, or *allow*, but also future intent and certain speech acts (e.g. *ask*) (see Quirk et al. (1985) for a full list). In combination with *that* verb complements, we interpret them as markers of argumentative discourse with the aim to promote ideas, make an argument, or persuade an audience, as in Example (2-b). A look at the distribution of registers confirms this interpretation. Research articles are the most representative register on this factor, followed by FAQ and comments. Factor 2 is thus dubbed Dimension 2 "Informational-argumentative".

(2) a. These are issues of jurisdiction, however$_{\text{conjunct}}$, not privacy. (comments, comments_28791923 _54_MAT).

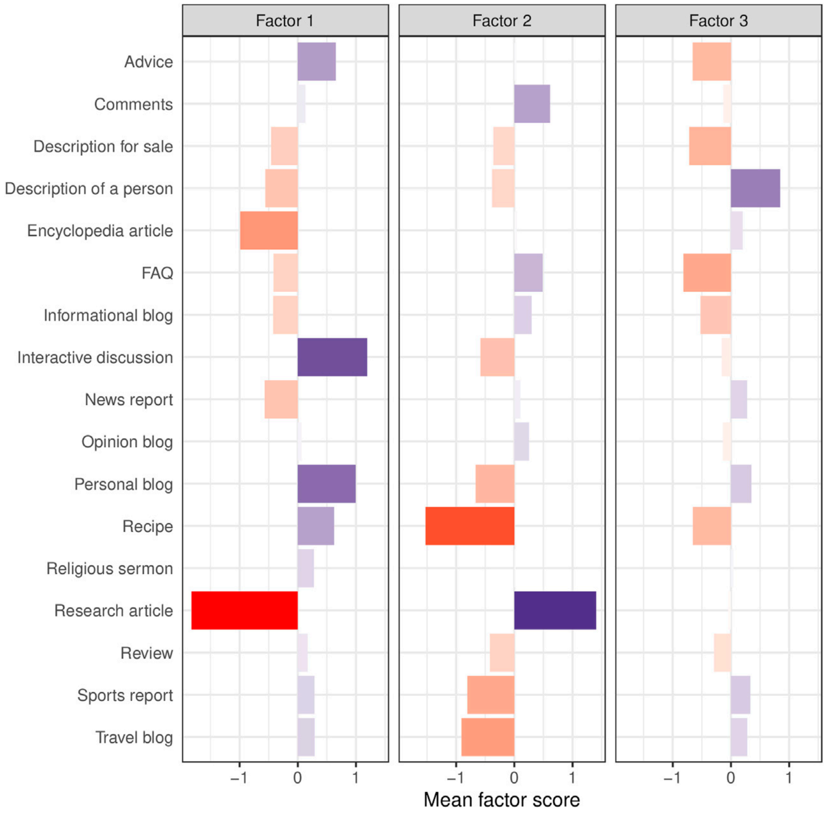


FIGURE 2 | Register distribution across the three factors/dimensions. Colour intensity indicates strength of mean factor scores. Red bars indicate negative values; blue bars indicate positive values.

b. He proposes$_{\text{suasive verb}}$ that$_{\text{that verb complement}}$ an individual might be genetically predetermined [. . .] (research article, 31_IDE_res_0026415_MAT.txt).

Factor 3 counts only four features with loadings ≥ |0.3|, and can, strictly speaking, not be fully and reliably linguistically interpreted. The interpretation provided here is therefore a tentative one, but we believe it is useful, as it supports the interpretation of the first two factors. Past tense, third person pronouns, and public verbs load on the positive pole of Factor 3 and are clear indicators for a narrative style (Biber, 1988, 108). Furthermore, the non-salient feature *that* deletion with a loading of 0.264 suggests description or elaboration of information—although this feature is common in spontaneous production (Biber, 1988, 244). Representative registers on the positive pole of Factor 3 are description of a person, personal blog, and sports report. Such registers describe or narrate events, actions, or people in a spontaneous or informal fashion and thus correspond to the co-occurrence of the positive features described above.

There is only one salient negative feature on Factor 3: present tense. According to the literature, present tense usually occurs in spontaneous and involved discourse. To derive at a more dependable interpretation, we examine secondary, non-salient features with loadings | ≥ 2| which do not load higher with the same polarity on another factor. These features consist of second person pronouns (−0.26) and modals expressing possibility (−0.206). Together with present tense verbs, they can serve to convey instruction, direction, or advice as illustrated in Example 3. As a matter of fact, the most characteristic registers on the negative pole of Factor 3 are FAQ, description for sale, advice, and recipe. Factor 3 is thus tentatively labelled as Dimension 3 "Narrative-descriptive vs. instructional".

(3) a. If the feta is$_{\text{present tense}}$ more salty than sharp, you$_{\text{2nd person pronoun}}$ may$_{\text{possibility modal}}$ want to squeeze over a little lemon juice (recipe, 07_I_recipe_1478719_MAT.txt).

b. If you$_{\text{2nd person pronoun}}$ 're$_{\text{present tense}}$ expecting some kind of fairy tale ending, you$_{\text{2nd person pronoun}}$ can$_{\text{possibility modal}}$ forget$_{\text{present tense}}$ about that right now. (description for sale, 16_IP_sale_0010352 _MAT.txt).

All together, these three factors paint a clear picture of the nature of online comments and online registers. We find an involved vs. informational divide, a result that has consistently been found in multi-dimensional analyses to be a feature of most registers, including cross-linguistically (Biber, 1995), and thus proposed as a universal of register variation (Biber, 2014). In our case, that first dimension is also imbued with evaluative meaning, conveyed by *be* as a main verb and predicative adjectives, which is why we have characterised that Factor as "Involved-evaluative".

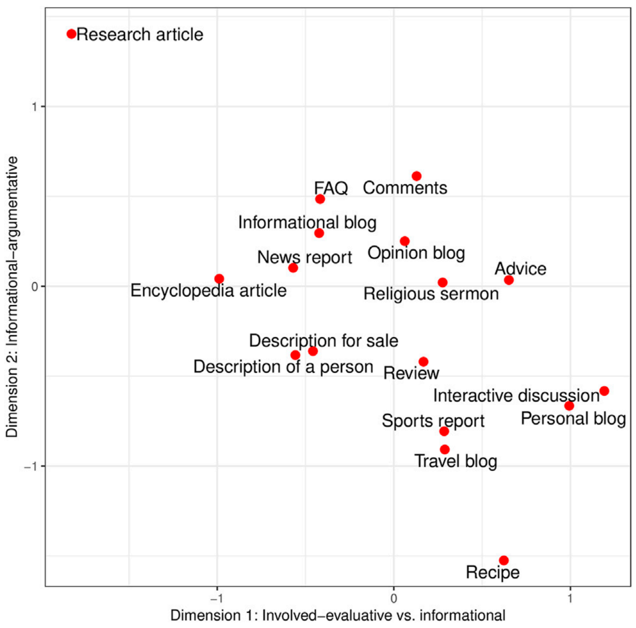


FIGURE 3 | Two-dimensional view of register distribution along the most prominent dimensions. Note that negative values on Dimension 1 represent the informational pole while positive values on Dimension 1 represent the involved-evaluative pole.

The "Informal argumentation" label for Factor 2 will be familiar to anyone who has spent any time online. One is likely to encounter vast amounts of argumentation, often involving a passionate defence of somebody's choice of movie, book, video game, or other artistic productions and consumer products. Argumentation, of course, is often deployed to defend or attack political ideas, argue for and against the conspiracy theory *du jour*, or to praise and vilify public figures. The web is an opinionated space and comments on news even more so. This is what Tufekci (2008) has described as the expressive internet.

Finally, Factor 3 points to the helpful and friendly aspects of the internet, a place where we can encounter descriptions and personal narratives, together with extremely helpful advice on the most esoteric or mundane aspects of everyday life, the instrumental internet (Tufekci, 2008). We can personally attest to the usefulness of the internet's collective wisdom when it comes to answering programming questions, solving plumbing issues, or fixing a bike. This factor combines that friendly aspect together with the construction of certain social personas (the helpful advice-giver, for instance).

4 ONLINE NEWS COMMENTS COMPARED TO OTHER ONLINE REGISTERS

After having interpreted the various factors as dimensions of variation, we will now turn to discussing the position of online news comments on the three dimensions relative to other online registers. **Figure 3** provides a two-dimensional view of the analysed registers along the major dimensions: Dimension 1 "Involved-evaluative" on the x-axis and Dimension 2 "Informational-argumentative" on the y-axis. For strength and direction of mean factor scores and register distribution, see also **Figure 2**.

On Dimension 1, online news comments (mean factor score 0.129) are positioned on the positive pole, i.e., they are mainly characterised by the joint occurrence of involved and to some extent evaluative features such as contractions, first and second person pronouns, adverbs, predicative adjectives, and *be* as main verb. However, in comparison to other online registers, online news comments exhibit comparatively few of these features. Registers such as interactive discussion, personal blog, advice, and recipe, for instance, are much more involved in nature than online comments. Thus, while online comments are positioned on the positive pole of Dimension 1, they also contain a fair amount of informational-abstract features such as average word length, nouns and nominalisations, prepositions and attributive adjectives—this can also be seen from their location on Dimension 2 (see below). The registers most closely positioned or similar to online news comments on the positive pole of Dimension 1 are review (mean factor score 0.168) and opinion blog (mean factor score 0.062). On the negative pole of Dimension 1, the most similar registers to the comments are FAQ (mean factor score −0.417) and informational blog (mean factor score −0.423). Titak and Roberson (2013) also found that reader comments were on the personal narrative pole, closer to e-mail and blogs, rather than on the informational pole.

Dimension 2 "Informational-argumentative" is the most characterising dimension for online news comments in this analysis: with a mean factor score of 0.613, they are one of the most representative registers on Dimension 2. They are clearly marked by the co-occurrence of nominalisations, a high average word length, conjuncts, *that* verb complements, and suasive verbs. As already mentioned in the previous section, all of these features contribute to creating informational and argumentative discourse. The other registers which are most representative of Dimension 2 are research articles (mean factor score 1.404) and FAQ (mean factor score 0.485)—both highly information-focused registers with an argumentative structure. The registers closest, and therefore most similar, to online news comments on this dimension are FAQ and informational blog (mean factor score 0.296), both marked by an informational-argumentative style, even though to a lesser extent than online news comments.

In regard to Dimension 3 (we remind the reader that the interpretation of this dimension is not conclusive) online news comments are rather instructional than narrative-descriptive. That said, their mean factor score on Dimension 3 is close to zero, which means that neither the features on the negative pole nor the features on the positive pole of this dimension are highly characteristic of online news comments. Typical instructional registers in this dataset are FAQ, description for sale, advice, and recipe. These registers are marked by a large amount of present tense forms and, to a lesser extent, second person pronouns and possibility modals. Registers representative of the narrative-descriptive pole are description of a person, personal blog, and sports report, which are marked by the co-occurrence of past tense verbs, third person pronouns, and public verbs. The registers most similar to online news comments (mean factor score −0.127) are research articles (mean factor score −0.045) and opinion blog (mean factor score −0.135) on the negative pole of Dimension 3, while the closest registers on the positive pole are religious sermon (mean factor score 0.044) and encyclopedia article (mean factor score 0.201).

According to their location on the three dimensions of variation, online news comments can best be characterised as instances of informational-argumentative discourse with a slight involved-evaluative slant. Anyone with experience reading online news comments will find this description apt: They tend to range from the preachy to the encyclopedic, with heavy argumentation. This characterisation is certainly intuitive if we consider the situational context in which online news comments are produced. Online news comments invite users to communicate their opinion on current news issues and can therefore contain involved and evaluative features (as indicated by their position on Dimension 1). However, online news comments are not subject to on-line production constraints and can be revised before posting, so that information can be integrated and commenters can make precise lexical choices to make their arguments (as indicated by their position on Dimension 2). This description is also in line with our other recent analyses. Ehret and Taboada (2020) compared online news comments to traditional written and spoken registers and found that they are strongly evaluative in nature, combining argumentative, informational, and some involved features (Ehret and Taboada, 2020, 23), while Cavasso and Taboada (2021) observe their overwhelmingly negative nature, with personal affective opinion (*I hate the candidate*) eschewed in favour of more detached evaluation (*The candidate is incompetent; The candidate's policies are bad*). As illustrated in (4), online news comments can thus range from involved-evaluative to involved-argumentative and informational-argumentative. In our analysis of exclusively written online registers, however, online news comments are not as prominently evaluative as other online registers and their evaluative nature did not emerge as a separate dimension of variation.

(4) a. $\text{I}_{\text{1st person pronoun}}$'$\text{m}_{\text{contraction}}$ $\text{very}_{\text{amplifier}}$ flattered that my writing $\text{is}_{\text{be main verb}}$ $\text{so}_{\text{emphatic}}$ $\text{powerful}_{\text{predicative adjective}}$ it scares $\text{you}_{\text{2nd person pronoun}}$. (comments, comments_3345 0158_18 _MAT.txt).

b. $\text{I}_{\text{1st person pronoun}}$ $\text{agree}_{\text{public verb}}$ $\text{that}_{\text{that verb complement}}$ more controlled peer reviewed research still needs to be done but let's$_{\text{contraction}}$ not run around $\text{saying}_{\text{public verb}}$ $\text{that}_{\text{that verb complement}}$ there is 0 scientific evidence. (comments, comments_7018634_53_MAT.txt)

c. $\text{However}_{\text{conjunct}}$, the SCC $\text{quite}_{\text{adverb}}$ $\text{often}_{\text{adverb}}$ throws back $\text{legislation}_{\text{nominalisation}}$ to the $\text{government}_{\text{nominalisation}}$ to redraft or abolish. (comments, comments_2463 0480_7_MAT.txt)

A large body of literature has explored the abusive and toxic nature of much online content and news comments in particular (McGuire, 2015; Gardiner et al., 2016; Muddiman and Stroud, 2017; Wolfgang, 2018; Juarez Miro, 2020). We found some toxicity in the comments in our corpus (Kolhatkar and Taboada, 2017a; Gautam and Taboada, 2019; Kolhatkar et al., 2020), but a relatively small amount, likely because the newspaper uses both automatic and human moderation to filter out the worst abuse.

Our previous analyses compared online news comments to other traditional registers (Ehret and Taboada, 2020), showing that they are not conversational at all. Here, we explore online registers in general and find that the nature of online registers is quite different from traditional written and spoken registers, and that comments are unique in the space of online registers. On the one hand, online registers are substantially more evaluative than traditional written registers—hence, online news comments do not emerge as strongly evaluative in this analysis. Although the fundamental distinction between involved and informational discourse (Biber, 1988) is still present in online registers, the scale of this continuum differs from analyses of purely traditional registers. On the other hand, online registers—and therefore also the emerging dimensions presented in this paper—seem not as clearly delineated as traditional registers in that they tend to combine features customarily associated with several (traditional) registers, and/or written and spoken language (Biber et al., 2015; Egbert et al., 2015). They are involved, like spoken language, but informational and argumentative like many written registers.

Our results contribute to the growing body of evidence that online registers are a different form of communication, and not a hybrid mode somewhere between speech and writing. Studies of Twitter (Clarke and Grieve, 2019), Reddit (Liimatta, 2019), and other online platforms (Hardy and Friginal, 2012; Titak and Roberson, 2013; Pavalanathan et al., 2017; Berber Sardinha, 2018), point to a new type of communication, including individual variation within the various platforms and communication channels. For instance, Liimatta (2019) found the now-familiar informational style in Reddit posts, but also, similar to the present analysis, an instructional focus. Berber Sardinha (2018) discovered two different types of stance (evidentiality and affect) in a study of a mix of online registers. Titak and Roberson (2013) placed reader comments in a personal narrative space (with orientation to the past) and also found that they tend to be involved. Hardy and Friginal (2012) found, like us and most other MDA studies, an informational vs. involved dimension in their analysis of blogs. Unlike the present paper, and due to the personal and narrative style of blogs, they additionally found addressee focus and narrative style dimensions. This makes perfect sense, as each platform and communication medium serves different communicative purposes, has different affordances, and is built around different communities of practice. Thus, the online space can be best described as a "continuous space of register variation" (Biber and Egbert, 2018, 196).

We should point out, before concluding, that our study is firmly language-dependent. The two corpora analysed are in English and it is quite possible that other languages may differ in the dimensions exhibited by different types of online registers. Biber (1995) shows that the main dimensions are constant across languages, especially the first dimension repeatedly found in multi-dimensional analyses (involved vs. informational). That result applies, however, to traditional written and spoken registers and may not be as robust in the online context.

5 CONCLUSIONS

This paper presented an analysis of online news comments in the context of other online registers. In particular, we conducted an MDA analysis to explore the linguistic features of online news comments compared to an extensive set of common online registers such as personal blog, advice pieces, or reviews.

Describing the position of online news comments along the three emerging dimensions, "Involved-evaluative", "Informational-argumentative", and "Narrative-descriptive vs. instructional", our results corroborate previous research on online news comments. A recent publication established online news comments as a separate register strongly different from other traditional written and spoken registers and described them as argumentative and evaluative instances of discourse (Ehret and Taboada, 2020). Although in the present analysis online news comments also turned out to combine an argumentative-informational style with some involved-evaluative characteristics, we found that online news comments are by far not as involved and evaluative as other online registers.

The analysis presented here thus further refines the previous description of online news comments and allows two general conclusions: First, online registers are not as clearly defined as traditional registers, because they combine features typically found in spoken and informal language with features typical of writing and formal language as well as feature combinations from multiple registers. Second, online registers tend to be more involved and evaluative than traditional registers. Although some online registers have consistently been shown to be involved (e.g. personal blog, advice) vs. other, more informational registers (e.g. research articles, informational blog), it is the involved plus evaluative makeup of online registers which marks them as distinct from other (traditional) registers. This unique combination of evaluative or opinionated features with informational, narrative, and descriptive styles has been previously noted and contributes to the hybrid nature of online registers (Biber and Egbert, 2016; Biber et al., 2015, for hybridisation of online registers see also; Santini, 2007).

These two general characteristics, their unique mix of spoken and written features combined with the involved-evaluative characteristics, suggest online registers are a different mode of communication, neither spoken nor written, and certainly not somewhere in the middle.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/sfu-discourse-lab/SOCC. Code for the study: https://github.com/sfu-discourse-lab/MDA-OnlineRegisters.

AUTHOR CONTRIBUTIONS

Following the CRediT system.[5] KE: conceptualization, statistical analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing. MT: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, validation, writing—review and editing.

FUNDING

The first author gratefully acknowledges funding from the Alexander von Humboldt Foundation through a Feodor Lynen Postdoctoral Research Fellowship. This research was also supported by the Social Sciences and Humanities Research Council of Canada (Insight Grant No. 435-2014-0171 to M. Taboada) and by NVIDIA Corporation, with the donation of a Titan Xp GPU.

The funder was not involved in the study design, collection,. analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

ACKNOWLEDGMENTS

Thanks to the members of the Discourse Processing Lab at Simon Fraser University, and especially to Laurens Bosman for help with data manipulation and feature extraction. Thank you to Jesse Egbert for sharing the CORE data. We are also grateful for feedback from the audience at the 4th Conference of the American Pragmatics Association in Albany (November 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.643770/full#supplementary-material

[5] https://casrai.org/credit/

REFERENCES

Berber Sardinha, T. (2018). Dimensions of Variation across Internet Registers. *International Journal of Corpus Linguistics* 23, 125–157. doi:10.1075/ijcl.15026.ber

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., and Davies, M. (2015). Exploring the Composition of the Searchable Web: a Corpus-Based Taxonomy of Web Registers. *Corpora* 10, 11–45. doi:10.3366/cor.2015.0065

Biber, D., and Egbert, J. (2016). Register Variation on the Searchable Web. *J. English Linguistics* 44, 95–137. doi:10.1177/0075424216628955

Biber, D., and Egbert, J. (2018). *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, D., and Finegan, E. (1989). Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect. *Text* 9, 93–124. doi:10.1515/text.1.1989.9.1.93

Biber, D. (2014). Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation. *LiC* 14, 7–34. doi:10.1075/lic.14.1.02bib

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Cavasso, L., and Taboada, M. (2021). A Corpus Analysis of Online News Comments Using the Appraisal Framework. *J. Corpora Discourse Stud.* 4, 1–38. doi:10.18573/jcads.61

Clarke, I., and Grieve, J. (2019). Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018. *PLoS ONE* 14, e0222062. doi:10.1371/journal.pone.0222062

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. New York: Routledge.

Dziuban, C. D., and Shirkey, E. C. (1974). When Is a Correlation Matrix Appropriate for Factor Analysis? Some Decision Rules. *Psychol. Bull.* 81, 358–361. doi:10.1037/h0036316

Egbert, J., Biber, D., and Davies, M. (2015). Developing a Bottom-Up, User-Based Method of Web Register Classification. *J. Assn. Inf. Sci. Tec.* 66, 1817–1831. doi:10.1002/asi.23308

Ehret, K., and Taboada, M. (2020). Are Online News Comments like Face-To-Face Conversation? *Register Studies* 2, 1–36. doi:10.1075/rs.19012.ehr

Farina, M. (2018). *Facebook and Conversation Analysis: The Structure and Organization of Comment Threads*. London: Bloomsbury Publishing.

Frehner, C. (2008). *Email, SMS, MMS: The Linguistic Creativity of Asynchronous Discourse in the New media age*. Bern: Peter Lang.

Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., and Ulmanu, M. (2016). The Guardian. The Dark Side of Guardian Comments. April 12, 2016. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

Gautam, V., and Taboada, M. (2019). Hey, Tyee Commenters! Scholars Studied You. Here's What They Found. Tyee. November 6, 2019. https://thetyee.ca/Culture/2019/11/06/Tyee-Commenters-Assessed/

Giltrow, J., and Stein, D. (2009). *Genres in the Internet: Issues in the Theory of Genre*. Amsterdam: John Benjamins.

Godes, D., and Mayzlin, D. (2004). Using Online Conversations to Study Word-Of-Mouth Communication. *Marketing Sci.* 23, 545–560. doi:10.1287/mksc.1040.0071

Halliday, M. A., and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hardy, J. A., and Friginal, E. (2012). Filipino and American Online Communication and Linguistic Variation. *World Englishes* 31, 143–161. doi:10.1111/j.1467-971x.2011.01728.x

Herring, S., Scheidt, L. A., Bonus, S., and Wright, E. (2004). Bridging the gap: A Genre Analysis of Weblogs. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences. Hawaii, 11.

Hunston, S. (2011). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Juarez Miro, C. (2020). The Comment gap: Affective Publics and Gatekeeping in the New York Times' Comment Sections. *Journalism* 1464884920933754.

Kolhatkar, V., and Taboada, M. (2017a). Constructive Language in News Comments. In Proceedings of the First Workshop on Abusive Language Online. Vancouver, 11–17.

Kolhatkar, V., and Taboada, M. (2017b). Using the New York Times Picks to Identify Constructive Comments. In Proceedings of the Workshop on Natural Language Processing and Journalism, Conference on Empirical Methods in NLP. Copenhaguen, 100–105.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2020). The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. *Corpus Pragmatics* 4, 155–190. doi:10.1007/s41701-019-00065-w

Liimatta, A. (2019). Exploring Register Variation on Reddit. *Register Studies* 1, 269–295. doi:10.1075/rs.18005.lii

Marcoccia, M. (2004). On-line Polylogues: Conversation Structure and Participation Framework in Internet Newsgroups. *J. Pragmatics* 36, 115–145. doi:10.1016/s0378-2166(03)00038-9

McCulloch, G. (2020). *Because Internet: Understanding the New Rules of Language*. New York: Riverhead Books.

McGuire, J. (2015). Uncivil Dialogue: Commenting and Stories about Indigenous People. CBC News, *November* 30, 2015. https://www.cbc.ca/newsblogs/community/editorsblog/2015/11/uncivil-dialogue-commenting-and-stories-about-indigenous-people.html

McVeigh, J. (2020). Thanks for Subscribing! A Genre Analysis of Email Marketing. *Language@ Internet* 18, urn:nbn:de:0009–7–51765

Muddiman, A., and Stroud, N. J. (2017). News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. *J. Commun.* 67, 586–609. doi:10.1111/jcom.12312

Nini, A. (2019). "The Multi-Dimensional Analysis Tagger," *Multi-Dimensional Analysis: Research Methods and Current Issues*. Editors T. Berber Sardinha and M. Pinto Veirano (London: Bloomsbury), 67–94.

North, S. (2007). 'The Voices, the Voices': Creativity in Online Conversation. *Appl. Linguistics* 28, 538–555. doi:10.1093/applin/amm042

Page, R. E., Barton, D., Unger, J. W., and Zappavigna, M. (2014). *Researching Language and Social Media: A Student Guide*. New York: Routledge.

Pavalanathan, U., Fitzpatrick, J., Kiesling, S., and Eisenstein, J. (2017). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, 884–895. A Multidimensional Lexicon for Interpersonal Stancetaking.

Peterson, E. E. (2011). How Conversational Are Weblogs? *Language@ Internet* 8, 2011 urn:nbn:de:0009–7–31201.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Santini, M. (2007). Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40). Hawaii: IEEE, 71.

Taboada, M. (2011). Stages in an Online Review Genre. *Text and Talk* 31, 247–269. doi:10.1515/text.2011.011

Titak, A., and Roberson, A. (2013). Dimensions of Web Registers: An Exploratory Multi-Dimensional Comparison. *Corpora* 8, 235–260. doi:10.3366/cor.2013.0042

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-rich Part-Of-Speech Tagging with a Cyclic Dependency Network. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Edmonton, 252–259.

Tseronis, A. (2011). From Connectives to Argumentative Markers: A Quest for Markers of Argumentative Moves and of Related Aspects of Argumentative Discourse. *Argumentation* 25, 427–447. doi:10.1007/s10503-011-9215-x

Tufekci, Z. (2008). Grooming, Gossip, Facebook and MySpace. *Inf. Commun. Soc.* 11 (4), 544–564. doi:10.1080/13691180801999050

van Eemeren, F. H., Houtlosser, P., and Henkemans, A. F. S. (2007). *Argumentative Indicators in Discourse: A Pragma-Dialectical Study*. Berlin: Springer.

Vásquez, C. (2014). *The Discourse Of Online Consumer Reviews*. London: Bloomsbury.

West, L. E. (2013). Facebook Sharing: A Sociolinguistic Analysis of Computer-Mediated Storytelling. *Discourse, Context & Media* 2, 1–13. doi:10.1016/j.dcm.2012.12.002

White, L. (2003). *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

Wolfgang, J. D. (2018). Cleaning up the "Fetid Swamp". *Digital Journalism* 6, 21–40. doi:10.1080/21670811.2017.1343090

Woollaston, V. (2013). Online Conversations Are Damaging How We Speak to Each Other in Real Life: Author Claims People Could Soon 'forget' How to Handle Social Situations. Daily Mail, September 30, 2013.

Zappavigna, M. (2012). *Discourse of Twitter and Social media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from NVIDIA Corporation. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/frai.2021.654154

Dissemination Dynamics of Receding Words: A Diachronic Case Study of *Whom*

Axel Bohmann[1]*, *Martin Bohmann*[2,3] *and Lars Hinrichs*[4]

[1]*Englisches Seminar, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany,* [2]*Institute for Quantum Optics and Quantum Information - Vienna (IQOQI), Austrian Academy of Sciences, Vienna, Austria,* [3]*Vienna Center for Quantum Science and Technology (VCQ), Vienna, Austria,* [4]*Department of English, The University of Texas at Austin, Austin, TX, United States*

We explore the relationship between word dissemination and frequency change for a rapidly receding feature, the relativizer *whom*. The success of newly emerging words has been shown to correlate with high dissemination scores. However, the reverse—a correlation of lower dissemination scores with receding features—has not been investigated. Based on two established and two newly developed measures of word dissemination—across texts, linguistic environments, registers, and topics—we show that a general correlation between dissemination and frequency does not obtain in the case of *whom*. Different dissemination measures diverge from each other and show internally variable developments. These can, however, be explained with reference to the specific sociolinguistic history of *whom* over the past 300 years. Our findings suggest that the relationship between dissemination and word success is not static, but needs to be contextualized against different stages in individual words' life-cycles. Our study demonstrates the applicability of large-scale, quantitative measures to qualitatively informed sociolinguistic research.

Keywords: dissemination, sociolinguistics, receding features, whom, relativizers, register

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Joshua Waxman,
Yeshiva University, United States
Avi Shmidman,
Bar-Ilan University, Israel

Correspondence:
Axel Bohmann
axel.bohmann@anglistik.uni-freiburg.de

Specialty section:
This article was submitted to Language and Computation, a section of the journal Frontiers in Artificial Intelligence

Received: *15 January 2021*
Accepted: *14 June 2021*
Published: *29 June 2021*

Citation:
Bohmann A, Bohmann M and Hinrichs L (2021) Dissemination Dynamics of Receding Words: A Diachronic Case Study of Whom. *Front. Artif. Intell. 4:654154. doi: 10.3389/frai.2021.654154*

INTRODUCTION

The Sociolinguistics of Emergence and Attrition

Sociolinguistic research is predominantly concerned with the emergence and spread of linguistic innovations, but has paid less attention to the dynamics of receding features. The canonical S-curve pattern of linguistic change (Labov, 1994) proceeds along three idealized stages—barely perceptible incipient change, rapid frequency increase through incrementation, and establishment of the feature within the community—to a theoretical steady state. Feature dynamics beyond this point are less well-understood. Yet, sociolinguists stand to gain insight from attention to receding features. These are of interest in their own right as part of a community's repertoire, but also because systematic comparison of the dynamics involved in feature emergence and attrition can lead to a more comprehensive understanding of linguistic change in general.

The dynamics of lexical emergence have recently been addressed through large-scale computational-statistical methods. Grieve et al. (2017) develop a procedure to identify emerging words in a corpus of 8.9 billion Twitter messages, based on initially low frequency and a high increase in frequency over a given time period. In a follow-up study, Grieve (2018) predicts the further success of 54 emerging words identified in Grieve et al. (2017) as a function of word length, part-of-speech, underlying word-formation process, and novelty of the word's referent. The latter predictor is shown to be particularly relevant in determining the frequency development of innovative words, whereas part-of-speech does not appear to play a significant role.

A further important predictor of a word's success is its social dissemination, defined by Altmann et al. (2011) as the ratio between the number of social units (e.g. speakers or texts) in a sample that use the word and the expected number of social units using the word. This expected number is calculated under the assumption of random spread of the word across social units, given its relative frequency and each social unit's total word count. Altmann et al. (2011) and Altmann et al. (2013) find higher dissemination scores to be a strong predictor of a word's continued increase in frequency.

The notion of social dissemination has been taken up in Garley and Hockenmaier (2012) as well as in Stewart and Eisenstein (2018). In both of these studies, its predictive power is less evident, which may in part be attributed to the inclusion of proper nouns in Altmann et al. (2011). Usage of these may be more directly linked to social dynamics than usage of general innovations (Stewart and Eisenstein, 2018: 4368). Stewart and Eisenstein extend the concept of dissemination from the social to the linguistic context of words. They calculate linguistic dissemination based on a comparison between expected and observed unique trigram frequencies in which a given word occurs and show, on the basis of several statistical models, that this metric effectively predicts future frequency developments.

These large-scale, quantitative findings are conceptually related to recent work in a more qualitative perspective. Squires (2014) traces how one specific phrase coined by a TV personality is taken up on Twitter. After being used by fans of the show the phrase originates from in direct reference to the initial situation of utterance, the phrase gradually spreads to wider discursive contexts and becomes increasingly detached from its origin. Squires (2014) refers to this process as "indexical bleaching." Given that indexicality describes the connection of a sign to the specific contexts it is embedded in, the notion of indexical bleaching may be related to Altmann et al.'s (2011) concept of social dissemination, including its extension in Stewart and Eisenstein (2018): the further a linguistic unit is indexically bleached, the more evenly disseminated it can be expected to be. One important thing to note about Squires' research is that her focusing on an individual form allows her to trace in more detail the indexical dynamics involved in its spread. As such, her analysis is able to go beyond a static relationship between indexical focus and a word's successful spread. She concludes that "indexical strength catalyzes uptake, but indexical loss facilitates diffusion" (Squires, 2014: 58).

This observation implies that the role of dissemination (which we take to be inversely related to indexical strength) in predicting a form's future frequency development may assume different shapes at different stages of that form's life-cycle. Most of the studies cited above have restricted their focus to the rapid emergence of innovative words, and to predictions about their relatively short-term success. Altmann et al. (2013) also consider the development of established words over longer time periods, yet their focus remains on frequency increase. The extent to which the dynamics of receding forms, i.e. those that are firmly established in the language but decrease in frequency, mirrors those of emerging ones is currently not well understood.

We focus on one particular such form, the relativizer *whom*, in order to shed light on the question of how frequency decline interacts with dissemination during an extended phase of attrition. In addition to implementing Altmann et al.'s (2011) original measure and Stewart and Eisenstein's (2018) extension of it, we also address dissemination across registers and topics. This is done on the basis of a multi-dimensional analysis (Biber, 1988) and a topic model for the corpus under consideration. In contrast to Altmann et al.'s (2011) approach, focusing on text-level properties like register and topic enables us to treat the range of texts in our corpus not simply as distinct units, but to systematically relate them to one another in terms of their linguistic characteristics and discourse content. Tracing the association between a form and specific register contexts and topics is arguably a more immediate window into indexical focusing than simply quantifying its presence or absence in a number of texts which are conceived as otherwise undifferentiated units. Compared to Stewart and Eisenstein's (2018) measure, our newly developed dissemination indices relate to characteristics of the textual environment on the whole, instead of the immediate collocation behavior of a word.

A Rapidly Receding Word: *Whom*

Standard English allows for nine different devices to introduce relative clauses (RCs): *that, which, who, whose, whom, when, where, why*, and ZERO (that is, the absence of an overt element introducing a relative clause). Competition among these forms is in part governed by categorical rules, e.g. the fact that *that* is only permissible for introducing restrictive RCs, and in part by probabilistic constraints. The latter have been the focus of many recent studies and are relatively well-documented for the three most prolific members of the set, *which*, *that*, and *zero* (e.g. Guy and Bayley, 1995; Levey, 2006; Hinrichs et al., 2015). In addition to language-internal constraints like antecedent noun phrase length, RC length and whether the relativizer assumes the subject or object role in the RC, Hinrichs et al. (2015) show that relativizer choice is susceptible to the influence of prescriptivist norms. Together with broader stylistic drifts, such as the colloquialization of written English (Leech et al., 2009), these factors account for a marked frequency decrease of *which* during the second half of the 20th century.

Although characterized by a similarly drastic decline in frequency over the past 200 years, *whom* has received comparatively less attention. This form is commonly regarded as a case-marked variant of *who* expressing objective case, analogous to the correspondence between *she* and *her*, *he* and *him*, etc. (although see Lasnik and Sobin (2000) for a competing account). Accordingly, traditional prescriptive grammar would require *whom* instead of *who* in RCs with human antecedents in which the relativizer occurs in the object position (Aarts, 1994: 73), as in (1).

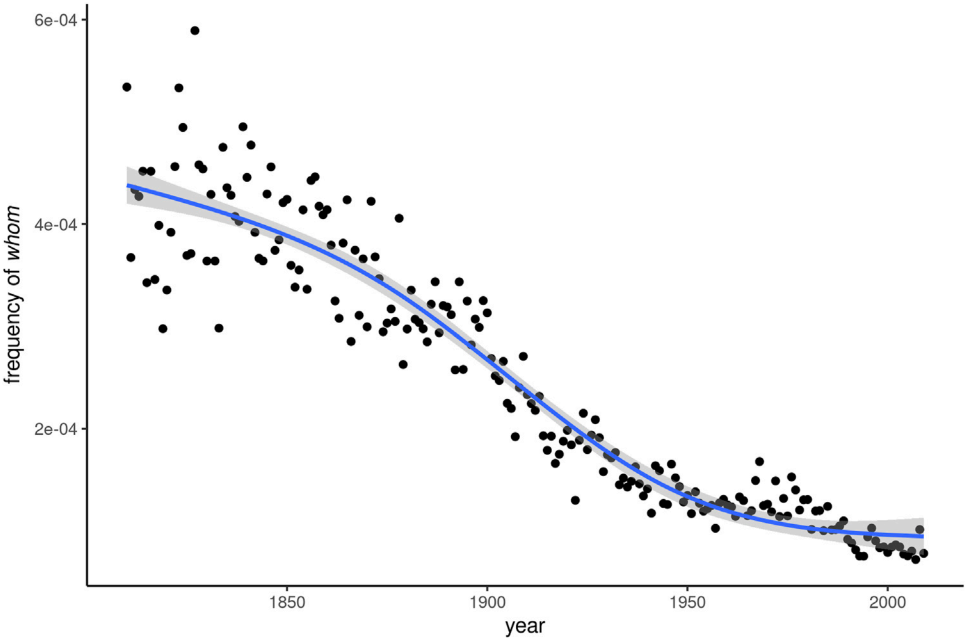


FIGURE 1 | Frequency development of *whom* over 180 years of written American English.

(1) going for the jugular of anyone **whom** he considers an enemy <COHA_fic_1988_782035>

As early as 1921, Sapir (1921: 167) predicted that "within a couple of hundred years from to-day not even the most learned jurist will be saying 'Whom did you see?'" Sapir identified several factors that conspire to render *whom* a moribund form: the general erosion of the English case-inflectional paradigm, the isolation of *whom* from the case-invariant remaining relativizers on the one hand and the system of personal pronouns on the other, as well as a purported "clumsiness" (Sapir, 1921: 171) in its phonetic shape. He further anticipated a general retreat of *who* and its variants from the class of relativizers in favor of highlighting their role as interrogative pronouns.

Many of these predictions have been borne out over the past 100 years. In the Corpus of Historical American English (COHA; Davies, 2010), the relative frequency of *whom* is consistently about an order of magnitude smaller than that of *who* throughout the 20th century. In terms of relative frequency, COHA shows a steady decrease of *whom* between 1810 and 2009, as can be seen in **Figure 1**. Using the Spearman correlation coefficient between relative frequency and year as an operationalization of the rise or fall of a word, *whom* is the fifth most rapidly receding item in the entire corpus, after *shall*, *nor*, *vain*, and *whence*. Along with this general decline, linguists have noted an increasing stylistic restriction to formal contexts, with prescriptivist discourse as an important catalyst (Aarts, 1994).

Figure 1, however, also shows that the rate of decline has slowed considerably in the second half of the 20th century. Empirical research on the recent past of written English has come to varied conclusions as to the fate of *whom*. Bauer (1994: 76) contends that avoidance of the word "has probably been noticeable throughout the [20th] century" and that ongoing change is relatively negligible, an observation also shared by Mair (2006: 141–143). Aarts and Aarts (2002: 128), on the other hand, find "staggering" rates of decline, both in written and spoken corpora, between the 1960s and the 1990s. **Figure 1**, based on larger and more systematic corpus data than the studies previously cited, would seem to confirm Mair (2006: 143) verdict that "*whom* now seems to have reached the tail end of the characteristic S-shaped curve of progression in linguistic change."

Despite disagreement about the most recent frequency developments, there is overwhelming consensus in the literature regarding the stylistic aspects of *whom*, namely: a strong association with very formal, almost exclusively written kinds of discourse. The fact that *whom* has not completely disappeared from the language is often attributed to its institutional backing in the educational system (Mair, 2006: 134). A discrepancy between actual usage and prescriptive norms means that most people "will recognize it as correct in a wider range of contexts [. . .], but probably not use it" (Bauer, 1994: 77).

The strong stylistic connotations of *whom* are evident in meta-linguistic discourse as well. In present-day internet culture, a class of memes is circulating which capitalizes on these indexicalities. The structural template for these memes pairs a sequence of images with a sequence of words. The images are repetitions of the same motif, a stylized X-ray of a human head, showing a rise in brain size with every iteration. The words form the sequence *who—whom—whoms—whomst'd*.[1] The rhetorical effect is an equation of linguistic forms with levels of intellectual

[1] See https://medium.com/write-i-must/dank-etymology-the-middle-english-origins-of-whomst-374ecd7a96fa for examples and an extended journalistic discussion.

superiority. The fact that both *whomst* and *whomst'd* are nonce words created for the context of this meme indicates the level of metalinguistic play inherent in it. These two words are constructed by attaching graphemic material to the base word that does not add any semantic content. In the case of *whomst*, it is likely that the *-st* sequence is used in analogy to archaic second-person singular verb inflections that were still common in Early Modern English. The grammatical information these suffixes used to bear is nowadays encoded on the subject only. The position of *whom* in this sequence construes this form as similarly burdened by unnecessary graphemic material but indicative of intellectual attainment. The meme consequently suggests a change in status for *whom* in that it has largely lost its grammatical function of case-distinction but gained indexical strength linking it to educated and hyper-formal contexts.

The properties described above make *whom* a suitable candidate for a contextualized analysis of various dissemination measures. Its frequency development over the past 200 years follows a clear trajectory which mirrors that of the S-curve often observed in the spread of linguistic innovations. The factors contributing to its decline, while not yet analyzed in a quantitative perspective, are well-attested. In addition, metalinguistic discourse surrounding the correct usage of *whom* in the form of prescriptive and descriptive linguists' comments is documented for at least as far back as the 18th century (Aarts, 1994). These facts enable us to formulate specific hypotheses regarding the dissemination of *whom* at different time periods and to contextualize observable dissemination developments against prior knowledge about the feature.

Research Objectives

We investigate the dynamics of dissemination that *whom* has undergone over the course of 180 years, between 1830 and 2009. Based on four quantitative measures, two established and two newly developed ones, we trace change in the dissemination of *whom* in this time period, which is characterized by continuous, but abating frequency decline. As can be seen in **Figure 1**, this decline is particularly rapid in the second half of the 19th and the first half of the 20th century, with the slope flattening again after around 1950.

On the basis of the literature on success during emergence, summarized in *The Sociolinguistics of Emergence and Attrition*, it would be valid to expect decrease in frequency to correspond with decrease in dissemination. This is the general statistical relationship that obtains in all the quantitative studies cited above, and is also a plausible hypothesis on purely theoretical terms. In the power-law distribution of any language's vocabulary, the most common items are likely shared by all speakers and across contexts, whereas low-frequency items in the long tail of the distribution can be expected to show stronger contextual sensitivity (Kretzschmar, 2015), i.e. lower dissemination. As a word's general frequency declines, one may consequently expect it to specialize into narrower niches of usage. In analogy to Squires' (2014) term, we call this process "indexical focusing." The tendency of receding forms to cluster in formulaic expressions serves as a case in point. In its extreme version, this process leaves receding words entirely unproductive and semantically intransparent outside of the larger constructions they are embedded in. Examples of such items include the highlighted words in the expressions *to make short shrift* or *kith and kin*. The baseline hypothesis for the analysis below, then, is that the frequency decline of *whom* will coincide with a decline in dissemination. However, Squires (2014) reminds us that this relationship may not be static.

Our focus on an individual word comes at the expense of generalizability. There is no guarantee that the dynamics we observe for *whom* are shared by all, or even the majority of, receding forms in the language. While recognizing this limitation, we suggest that this narrow focus also brings important advantages. In order to make statistical generalizations like those described in Altmann et al. (2011), Altmann et al. (2013), or Stewart and Eisenstein (2018) more immediately relevant to sociolinguistic research, they need to be understood in relation to individual features of interest. Unlike phenomena in statistical physics and other core sciences, words in a language are not merely units with certain statistical properties, but are embedded in individual histories of social meaning and metalinguistic reflection. The sociolinguistic record contains a large number of features about which a good deal is known in this respect. It is consequently possible to formulate specific expectations as to the relationship between frequency developments and dissemination measures for such features that go beyond general regularities. A consideration of individual words' social role in conjunction with observable statistical properties promises to enrich our understanding of both these perspectives.

Our aim is to make the notion of dissemination tangible from a situated sociolinguistic perspective and to evaluate the utility of each dissemination measure for future application in contextualized sociolinguistic research. Specifically, we ask how well the four measures correlate with change in frequency, as well as how strongly correlated they are with each other. If no correlation between frequency and a given dissemination measure can be found, the utility of that measure is up to question. If the dissemination measures show no or only weak correlation amongst each other, this fact requires further attention. Our assumption is that, despite being operationalized at different levels, dissemination is a general property which we expect to take a similar shape independent of its precise quantification.

MATERIALS AND METHODS

Corpus

Our analysis is based on the Corpus of Historical American English (COHA; Davies, 2010), which includes samples of written American English for each year between 1810 and 2009. The corpus is sub-divided into four genres: news, magazine, fiction, and non-fiction writing. Each word in the corpus is annotated with lemma and part-of-speech information.

Due to the difficulty of sampling historical language data, several aspects of the COHA sampling frame are not consistent throughout the 200 years it covers. For instance, the sparsity of texts for some genres from the more distant past has resulted in the inclusion of fewer, but longer individual texts for much of the 19th century. Further, newspaper texts are only sampled from 1860 onwards and different archives were used for the extraction of text samples for different time periods.[2] The effect of archival sources is visible especially for magazine writing, for which our register analysis (see below) shows a marked difference between texts before and after 1900.

Consideration of the above factors led us to exclude the first two decades of COHA (1810–1829) from the analysis. With a median number of 14.5 texts per year, these do not offer sufficient data for our analyses, most of which treat individual texts as the relevant units. We further note that the irregularities mentioned above are not fully resolved before the sampling point 1925. From this time on, both the archives used for text sampling and the mean number and word count of texts per year are consistent. While our analysis covers the years from 1830 up to 2009, then, the results are expected to be most robust for the latter half of this time period.

We work with the full-text, offline version of COHA, which includes lemma and part-of-speech information for each word. For each year between 1830 and 2009, we calculate the four dissemination measures for *whom* described in the following sections.

Social Dissemination

Following Altmann et al. (2011), we measure social dissemination (D^S), as the ratio between the observed and expected social units a word occurs in at a given time. For our purposes, the social units of relevance are the individual corpus texts. In other words, we divide the number of documents *whom* occurs in by the number of documents it is expected to occur in. To calculate the latter number, a probability of observing *whom* in each text is calculated based on the text's word count and the relative frequency of *whom* in the corpus at the time point under consideration. These probabilities are then summed to approximate the expected document count. The assumption for this baseline model is that all words occur randomly in the texts, with a probability corresponding to their relative frequency. The probability to find the word *whom* at least once in the i^{th} text of word length m_i is then given by $T_i = 1 - e^{-fm_i}$, where f is the relative frequency of *whom* in the considered year. Based on this, we can calculate the expected number texts containing *whom* via $\tilde{T} = \sum_{i=1}^{N_T} T_i$, where N_T is the number of texts in the considered year. With this expectation of the baseline model, we can calculate the dissemination coefficient

$$D^S = \frac{T}{\tilde{T}}$$

which is the ratio between the number of texts in which *whom* is used T and the expected number of texts following the baseline model. A value of $D^S = 1$ corresponds to dissemination of a word across texts as if its occurrence was entirely random. Values below 1 indicate "clumping" (Altmann et al., 2013: 3), i.e. the use of the word in a smaller set of texts than expected. The closer to 0 D^S is, the less regularly disseminated the corresponding word is. Under-dissemination is interpreted by Altmann et al. (2013) as a sign of low word vitality.

Linguistic Dissemination

Stewart and Eisenstein (2018) define linguistic dissemination (D^L) as the difference between the log count of unique trigrams a word occurs in (C^3) and the word's expected log unique trigram count ($\tilde{C}^3$). Since the logarithms of frequency and unique trigram count are highly correlated (Egghe, 2007; Stewart and Eisenstein, 2018: 4364), it is possible to calculate the expected log trigram count based on a word's frequency. In Stewart and Eisenstein (2018), this is done by fitting a linear model for all words at a given time point, with the words' log frequencies as the predictor and their log trigram counts as the outcome variable. Linguistic dissemination is then defined as the residual error between the model prediction and the observed log trigram count ($D^L = C^3 - \tilde{C}^3$). Positive values indicate higher-than-expected numbers of trigrams, i.e. particular linguistic versatility, whereas negative values indicate a restriction of the linguistic contexts a word occurs in. Negative D^L is a predictor of frequency decline.

We treat individual sentences as the relevant context for trigram detection and do not consider trigrams across sentence boundaries. Each document in the raw, unannotated version of COHA is split at sentence-final punctuation marks (periods, question and exclamation marks, semicolons, and colons). For copyright reasons, the offline version of COHA replaces sequences of ten words at set intervals with ten @ symbols. We treat these like sentence-final punctuation and do not allow trigrams to extend across them. If a word occurs in a place in the sentence that does not permit a right or a left trigram neighbor, i.e. in the first, second, last, or second-to-last position, we still register three unique trigrams. In these cases, we insert "<START>" or "<END>" instead of actual words into the trigram in order to replicate the method in Stewart and Eisenstein (2018).

Counting all trigrams for each word at a given time period proved computationally intractable. We therefore restrict ourselves to a random selection from a list of 17,912 words that occur at least 1,000 times in the corpus on the whole. For each time period, 10,000 items from this list of words are drawn and their unique trigram counts and frequencies of occurrence are measured. Given the regular relationship between log frequency and log unique trigram count, this amount of data is sufficient to reliably estimate the coefficient of the linear model and hence D^L.

Register Dissemination

In addition to social and linguistic dissemination, we also propose a measure of register dissemination (D^R). Our notion of register is closely in line with that developed by Biber (e.g. Biber, 1988; Biber and Conrad, 2009; Biber, 2012), both in how we conceptualize and how we quantify it. The term is defined as "a variety associated with a particular situation of use" (Biber and

[2] See https://www.english-corpora.org/coha/

Conrad, 2009: 6). While the relevant situational parameters may relate to medium and context of communication, communicative goals and norms, and a number of other extra-linguistic factors (see Biber and Conrad, 2009: chap. 2), they have a direct and measurable bearing on the linguistic properties of a stretch of discourse.

To measure the interrelationship between situational properties and linguistic characteristics, the exploratory method of multi-dimensional analysis (MDA; Biber, 1988) has been established in the corpus-linguistic community. This method proceeds by compiling a corpus of relevance for the analysis, i.e. one that represents the situational parameters of interest, as well as a number of linguistic features hypothesized to play an important role in register differentiation. Such features are usually relatively common, high- to mid-frequency ones, such as the frequency of passive-voice constructions, personal pronouns, or non-standard words in a text. For each corpus text, the frequency profile of each feature is measured. The resulting text-feature matrix is subjected to exploratory factor analysis (Thompson, 2004) in order to discover a small number of latent "dimensions of variation" (Biber, 1988) that capture a large amount of the total variance of the extracted features. Each dimension is characterized by the linguistic features it is most strongly associated with, and each corpus text can be scored on a continuum for each dimension. Qualitative consideration of the most strongly associated features and the highest- or lowest-scoring kinds of texts for a dimension drives the interpretation and labeling of each dimension.

We perform such an analysis for the entirety of the COHA data. We use 65 of the features proposed in Biber (1988) and 24 additional ones largely adapted from Bohmann (2019). In addition, we also include the relative frequency of each of the 100 most common part-of-speech trigrams in COHA. The resulting 116,614 × 179 text-feature matrix is subjected to factor analysis with the psych package (Revelle, 2020) in R (R Core Team, 2020). Following an inspection of the variances accounted for by the first 100 components of a principal component analysis over the features, we decided to extract five factors from the data. We use a principal axes factor solution rotated to the promax criterion, which allows for moderate inter-factor correlations. The factor scores for each text are calculated using the regression method (see Thompson, 2004; Revelle, 2020 for details).

Space does not permit a full discussion of the dimensions and the qualitative process that produced interpretations and labels for each. Here, we restrict ourselves to an overview in tabular form. **Table 1** shows the five dimensions (i.e. factors developed in the factor analysis) with the labels we have chosen for them. The most strongly associated features, genres, and the dimensions' development over time give an indication of what aspects of linguistic variation each captures.

Both the social and linguistic dissemination measures are based on discrete counts, which are not available for register as we operationalize it. A different method for quantifying register dissemination is therefore required than those used for social and linguistic dissemination above. Two options suggest themselves. First, similarly to Altmann et al. (2011) we can treat the presence or absence of *whom* in a text as a binary variable. For each step in the time period under analysis, we can then divide our corpus in two groups of texts, those including and not including *whom*. Both of these groups can be characterized as multivariate Gaussian distributions in the five-dimensional register-score space. Register dissemination can then be treated as the distinctiveness of the *whom*-texts from those without *whom* in register space. If there is significant overlap between both groups, this can be taken to indicate relatively wide dissemination, whereas if the groups are found to be largely distinct, this is a sign of register-specificity. The amount of overlap between two multivariate Gaussian distributions can be expressed as the Bhattacharyya distance (Bhattacharyya, 1943) between them.

This method is susceptible to differences in text length, since longer texts have a higher baseline probability of including a given feature and hence ending up in the *whom*-group. One solution would be to sub-divide larger texts into smaller segments to achieve uniform text length, and to treat each segment as a sample in its own right. While this would be a feasible solution in principle, a more plausible one is to treat relative frequency of *whom* in a text as a scalar variable. Doing so accounts for the effect of text length in a principled way without requiring further manipulation of the data. Instead of creating distinct groups, this method situates texts on a *whom*-frequency continuum.

In order to quantify the association between *whom* and specific register properties, we fit a linear model at each year, with relative frequency of *whom* as the outcome and each text's scores for the five dimensions as the predictor variables. The adjusted R^2 values of these models are taken as indices of register specificity. A dissemination coefficient with similar properties to that proposed by Altmann et al. (2011) can then be obtained by subtracting this adjusted R^2 from 1. The more predictive power the joint dimension scores yield regarding relative frequency of *whom*, the higher the model's R^2 value and the lower the corresponding D^R. As with Altmann et al.'s (2011) index, a value of 1 indicates completely even dissemination in register space, whereas values below 1 suggest register clumping and consequently decreased vitality of the form.

In addition to the general D^R, the values of each dimension's model coefficients can also be traced over time, giving a sense of which register dimensions are most predictive of *whom*-frequency and which are most subject to change over time.

Topic Dissemination

Apart from social and register properties, discourse topic may be an important predictor of linguistic variation. We create a topic model for COHA, which we restrict to 100,000 randomly selected texts for computational reasons. Specifically, we use latent Dirichlet allocation (LDA), which represents a predefined number of topics as probability distributions over the words in the corpus and treats every corpus text as a probability distribution over all topics (Blei et al., 2003).

Before generation of the topic model, the corpus data were preprocessed in the following manner: all words were lemmatized based on the information already included in COHA, and only words from the part-of-speech categories noun, verb, adjective,

TABLE 1 | The five dimensions of variation in COHA.

Dimension label	Most salient features	Genre differentiation	Diachronic development
Structural elaboration	Clausal coordination, noun phrase trigrams, prepositions, main verb BE, attributive adjectives	Highest in nonfiction writing; lowest in magazines and newspapers	Consistent decrease in all genres
Verbal-personal communication	Pro-verb DO, private (cognitive) verbs, verbal infinitives, first person pronouns, adverbs	Highest in fiction writing, lower in all other genres	Increase in the non-fiction genres in the second half of the 20th century
Information density	Attributive adjectives, mean word length, type-token ratio, nouns, prepositions	Highest in nonfiction, lowest in fiction	Increase in all genres, particularly in the 20th century
Narration	Simple past, third person pronouns, possessives, quotation marks, public (quotative) verbs	Highest in fiction, lowest in nonfiction	Consistent increase in fiction; irregular developments in other genres
Abstraction & generalization	Prepositions, nominalizations, agentless passives, mean word length, infinitives	Highest in newspapers and nonfiction writing, lowest in fiction and magazines	Newspaper, nonfiction, and magazine writing grow closer to the consistently low values for fiction

TABLE 2 | The 25 LDA topics developed for COHA.

Topic	Top words	Topic	Top words
1	candidate, democrat, kennedy, nixon, reporter	14	george, mexico, mexican, madeline, rollo
2	team, player, film, coach, movie	15	prince, queen, lord, rome, duke
3	boat, captain, sail, deck, crew	16	chinese, china, mountain, stone, surface
4	railroad, machine, steel, contract, profit	17	percent, budget, investment, oil, sales
5	soul, heaven, lord, dear, sir	18	peter, shot, int, sam, camera
6	p.-a., dear, aunt, mary, sir	19	sir, captain, colonel, horse, soldier
7	kid, guy, stare, phone, nod	20	patient, hospital, <br>, medical, drug
8	paul, bird, planet, flower, moon	21	senate, teacher, governor, amendment, candidate
9	cook, milk, fruit, sugar, meat	22	moral, religion, science, christian, religious
10	horse, wood, dog, mountain, stare	23	tom, joe, ben, ruth, phil
11	poet, poem, jane, poetry, novel	24	novel, magazine, editor, publisher, reader
12	animal, research, science, scientist, cell	25	governor, indian, lincoln, county, trial
13	soviet, russian, communist, germany, russia		

and adverb were retained. Sequences of proper nouns such as "United States" were treated as single words, once again drawing on the information provided in COHA. Finally, the top 1,000 bigrams and trigrams with a minimum absolute frequency of 100 were also treated as single units. Extraction and ranking of bi- and trigrams was done with NLTK's collocations module (Bird et al., 2009), which uses pointwise mutual information as its association metric.

The LDA models themselves were constructed in Python's (Python Software Foundation, 2020) gensim module (Řehůřek and Sojka, 2010), with the parameters chunksize set to 2,000, passes to 5, and iterations to 200. Such models were built for numbers of topics between 9 and 200. For each of these, model coherence was calculated with the C_v measure proposed in Röder et al. (2015). The candidate with the highest coherence is a 25-topic model. As with the register dimensions, it is not our aim here to discuss individual topics. Therefore, **Table 2** simply shows the top five words in each topic to give a sense of the range and plausibility of the model on the whole.

Our procedure of quantifying topic dissemination is largely the same as that for quantifying register dissemination, with one addition. The factor analytic procedure that produces register scores ensures that these are already uncorrelated, or only moderately correlated in the case of oblique rotation methods (Thompson, 2004). The opposite is the case for the topic probabilities of each text. Since these always sum to 1, they are fully collinear as a set and cannot be used directly as model predictors. We therefore subject them to principal component analysis and use the values of the principal components as predictors. This has disadvantages if one wishes to explore the effects of individual topics, but is entirely robust for evaluating the predictive power of the topic structure on the whole.

RESULTS

Social Dissemination

The development of social dissemination (D^S) of *whom* between 1830 and 2009 is shown in **Figure 2**. In addition to plotting dissemination scores per year, the figure shows the curve of predicted values based on a generalized additive model with a cubic spline and four knots.[3] The same curve formula is used for all plots below and was chosen to strike a compromise between being able to address nonlinear relationships in the data on the one hand and being relatively robust to noise on the other. In the case of social dissemination, the curve almost perfectly approaches a straight line, whose confidence intervals overlap

[3] To fit this curve, the following command was added to the plot objects in R: geom_smooth(method="gam", formula = y ~ s(x, bs = "cs", k = 4)).

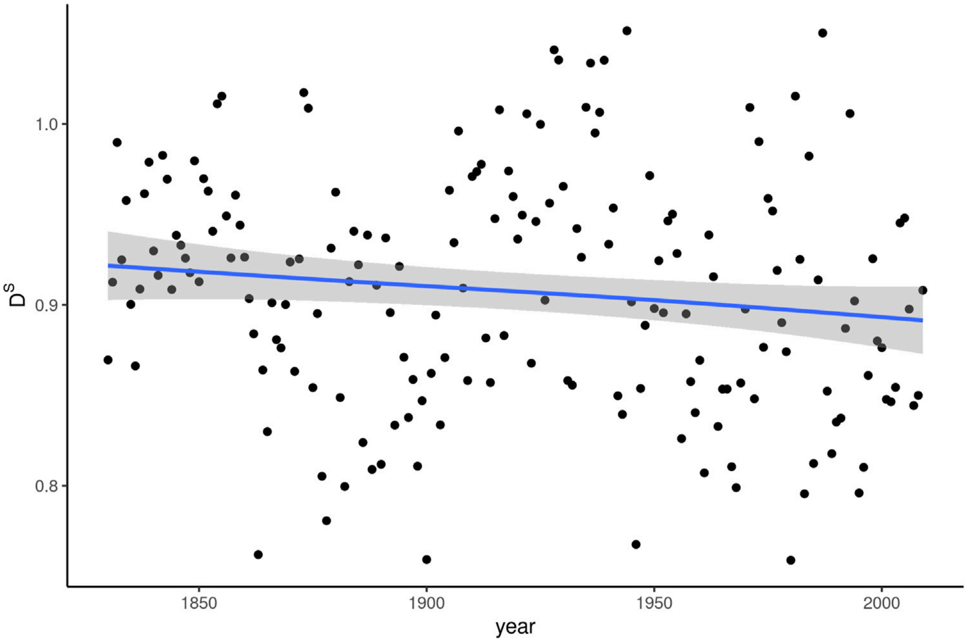


FIGURE 2 | Social dissemination (D^S) of *whom* over 180 years of written American English.

throughout the interval covered by the data. The slight negative slope is therefore of little statistical consequence. The considerable spread of individual, yearly D^S values throughout the period analyzed further confirms this impression.

The results shown in **Figure 2** are unspectacular on the whole. There is little information in this plot that sets *whom* apart from other words, either in terms of its general dissemination tendency or its dissemination dynamics over time. A mean social dissemination value of around 0.9 is entirely normal, since values below 1 are "in fact observed for most words" (Altmann et al., 2013: 3). *Whom* has a frequency of occurrence of about 2.5 * 10^{-4} in COHA. For words with similar frequency profiles, Altmann et al., (2011: 3) report median dissemination values around 0.8. The values in **Figure 2**, therefore, are above rather than below expected. Consequently, there is little indication that restricted dissemination accounts for the word's frequency decline.

Looking at the development of social dissemination over time, the stability of the values in **Figure 2** is moderately surprising. True enough, during the plotted time interval, *whom* continuously decreases in frequency. However, the decrease is not linear. It starts out relatively slowly, picks up speed in the second half of the 19th century and flattens out again after around 1950 (see **Figure 1**). Given that the relationship Altmann et al. (2011) and Altmann et al. (2013) establish is between change in frequency and change in dissemination, one would expect the bends in the frequency curve to coincide with changes in D^S, yet this is not the case.

We further tested for the correlation between Δf and ΔD^S, i.e. the change in both frequency and social dissemination measured in each year compared to the previous one. The result of a Pearson test did come out as significant ($p < 0.001$), but with a negative correlation of −0.261. This finding runs directly counter to expectations based on the attested relationship between frequency and social dissemination.

Linguistic Dissemination

Figure 3 shows the linguistic dissemination (D^L) indices along with the smoothing curve in the same fashion as **Figure 2**. The difference between the two plots is immediately apparent. D^L appears to undergo a much more dynamic development than D^S. The mean value of D^L for all words by definition is 0, with a standard deviation of 0.15 in our data. This means that *whom* starts out with linguistic dissemination values in the 19th century that are below average, but not strikingly so, as they are only about half a standard deviation from the mean. During the roughly 100 years of most intense frequency decline, between 1850 and 1950, linguistic dissemination actually increases steadily. In other words, while *whom* is receding in general, it appears to be gaining, not losing, linguistic versatility. In the second half of the 20th century, a time during which the frequency decrease slows down, D^L appears to reverse this upward trend.

The same correlation test as for the relationship between Δf and ΔD^S was also run for Δf and ΔD^L, with similar results. A coefficient of −0.387 ($p < 0.001$) confirms the impression from **Figure 3** that decline in frequency coincides with increase in linguistic dissemination.

In sum, linguistic dissemination develops almost entirely in the opposite direction from what might be expected based on the literature. Instead of a hypothesized positive correlation between D^L and frequency, extended periods of frequency decline coincide with a rise in D^L and periods of comparable frequency stability go hand in hand with a dip in linguistic dissemination.

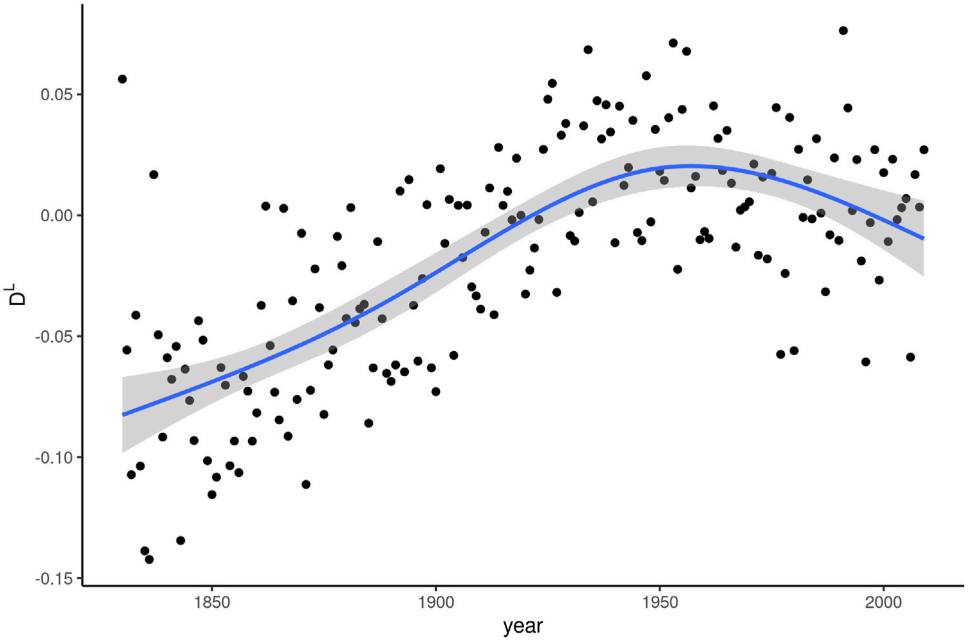


FIGURE 3 | Linguistic dissemination (D^L) of *whom* over 180 years of written American English.

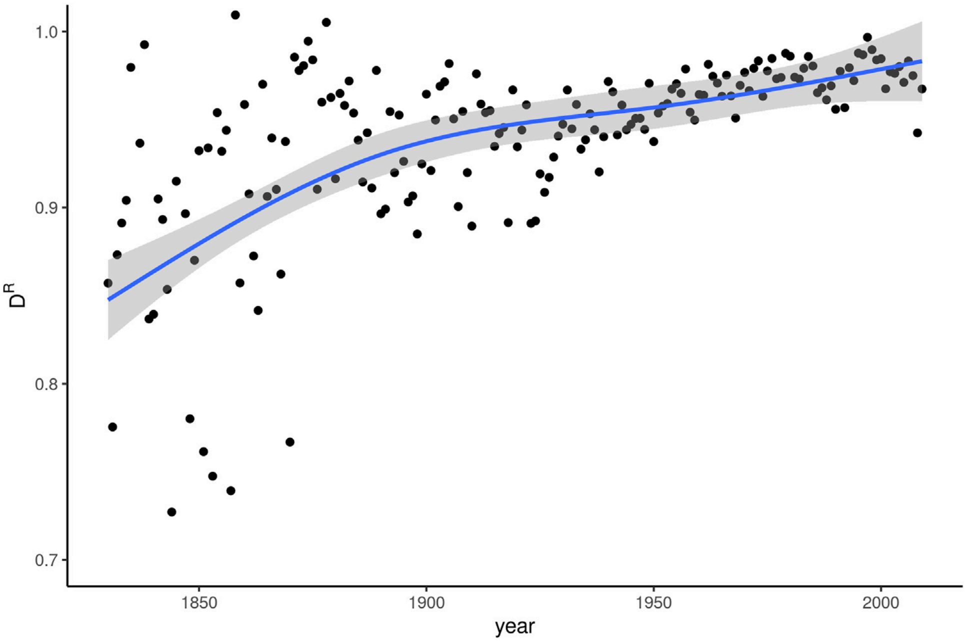


FIGURE 4 | Register dissemination (D^R) of *whom* over 180 years of written American English.

Register Dissemination

Figure 4 shows the D^R indices for linear models predicting relative frequency of *whom* based on the five dimension scores developed in the MDA described above. The curve of predicted values approaches a straight line with an upward slope, indicating that in earlier years, information about texts' dimension scores is better able to account for variation in *whom*-usage.

With the exception of the earliest decades in the data, D^R is consistently above 0.8. Accordingly, register plays only a very limited role in predicting the frequency of *whom*. The smoothing

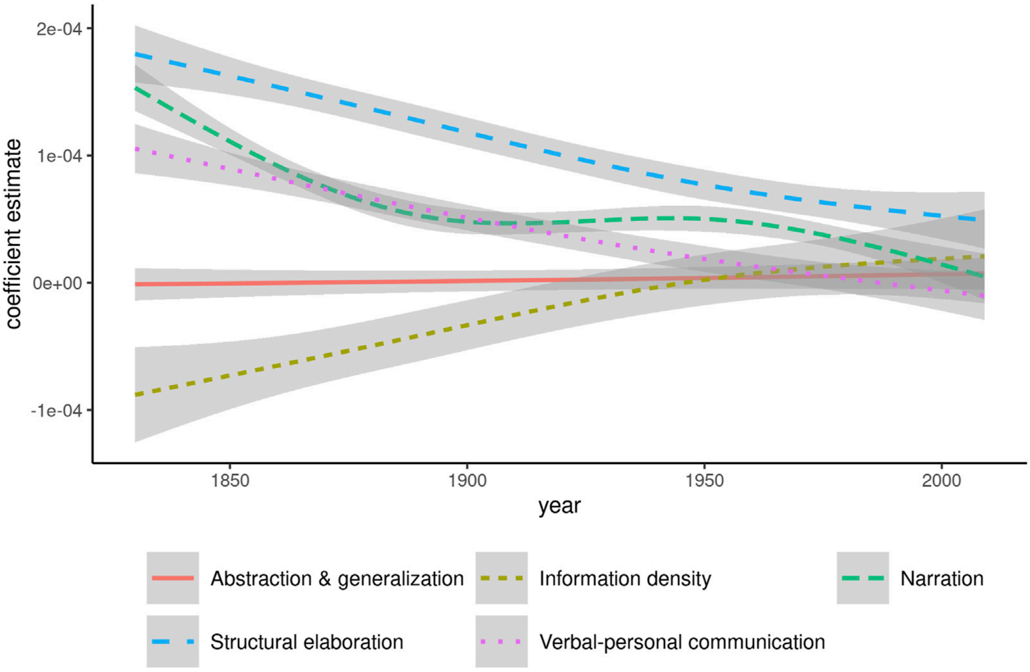


FIGURE 5 | Coefficients for individual dimensions in models predicting frequency of *whom*. Points for individual years were omitted to avoid overplotting.

curve approaches 1 for years after 2000. In the latest years of COHA, then, register information is almost entirely uninformative as to expected frequencies of *whom*. The conclusion has to be drawn that around the turn of the 21st century, there is almost no register differentiation left to characterize *whom* in actual usage. Once again, this pattern is directly opposite to expectations based on the known relationship between dissemination and frequency change for emerging words. A correlation test between Δf and ΔD^R produced no significant results. We assume that this is due to relatively large fluctuations in D^R for individual years and the comparatively longer time window over which register developments operate. Therefore, D^R may be better suited to address developments over more coarsely-grained time periods.

In relation to register, it is worth moving beyond the bird's eye view of all dimensions in conjunction and to consider how individual dimensions relate to the general trend identified in **Figure 4**. To this purpose, **Figure 5** shows smoothing curves of the coefficient estimates for each of the five dimensions calculated for each year.

NARRATION, STRUCTURAL ELABORATION, and VERBAL-PERSONAL COMMUNICATION all show a relatively steady regression from positive values towards 0. That is, an initial association between high text-scores along these dimensions and higher frequencies of *whom* decreases in strength in all three cases. The coefficients for ABSTRACTION & GENERALIZATION are indistinguishable from 0 throughout the entire period analyzed, showing that this dimension has no role to play in predicting the frequency of *whom*. INFORMATION DENSITY is the only dimension with an initially negative coefficient, which however increases steadily until it intersects 0 around 1950. From this point on, the coefficient values are positive, but so low that they are effectively indistinguishable from 0. This pattern suggests that *whom* is associated with comparatively loose packaging of information throughout the 19th century, but becomes increasingly associated with INFORMATION DENSITY in the 20th century.

At a more general level, the convergence towards 0, i.e. no measurable effect, for all dimensions is a striking pattern. By the 2000s, the only coefficient that is appreciably different from 0 is that for STRUCTURAL ELABORATION, and even this estimate is reduced to about half its value compared to the 1830s. **Figure 5** draws a much more vivid picture of the increasing register dissemination of *whom* throughout the period under analysis, a process that is almost complete by the last years covered in COHA.

Topic Dissemination

The story of topic dissemination is quickly told: no significant effect of discourse topic on *whom* can be discerned. This is readily apparent from **Figure 6**, which shows the D^T values derived from linear models predicting *whom* frequency based on (rotated) topic distributions in the texts. The values are very close to and at times even above 1. The latter is due to the models' R^2 being adjusted downward for predictor variables that add more complexity than predictive power, as is likely the case with some components derived from principal component analysis. More importantly, the minimal fluctuation in R^2 over time is not sufficient to indicate any diachronic pattern.

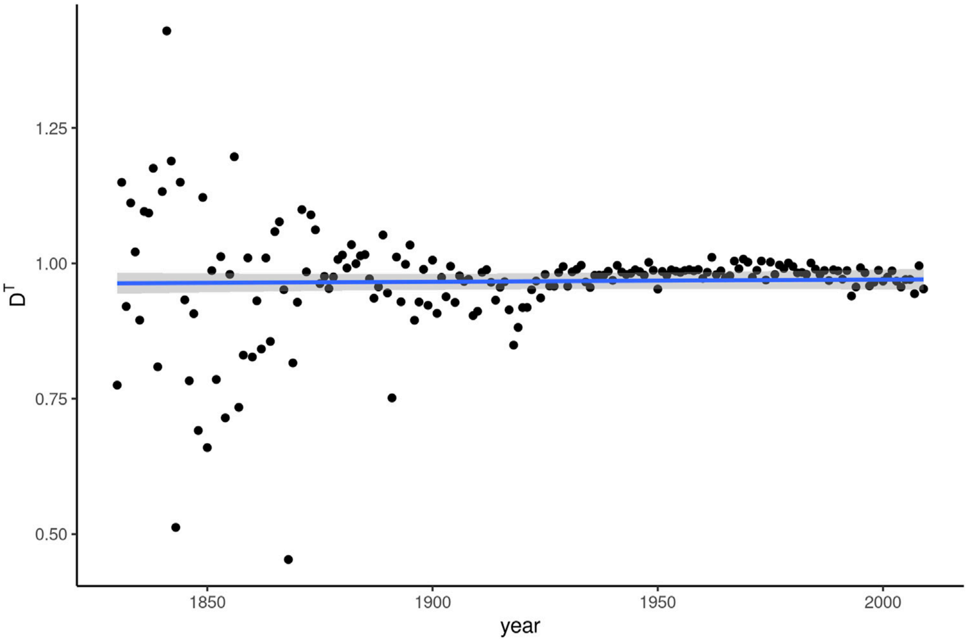


FIGURE 6 | Topic dissemination (D^T) of *whom* over 180 years of written American English.

DISCUSSION

Dynamics of Emerging Versus Receding Forms

Our results do not confirm any of the expectations one might derive from the extant literature on the relationship between word growth/decline and dissemination. The decrease in frequency of *whom* does not coincide with systematic decrease in any of the dissemination measures, nor do the individual measures themselves correlate to draw a unified picture. Pairwise Pearson tests between the four measures reveal one relatively weak correlation of note: linguistic and register dissemination show a Pearson's coefficient of 0.398 at $p < 0.001$. All other correlations do not reach statistical significance even at the least conservative conventional level of $p < 0.05$.

At least for the particular case of *whom*, then, there is little evidence to suggest that dissemination dynamics during the decline of an established feature parallel those during the emergence of innovative words. Results for a larger number of receding features are required to further substantiate the nature and degree of the differences suggested by the results above. We have explored the dissemination measures used here for the 200 most rapidly receding word forms in COHA and found item-specific differences to be more noticeable than any unified trend. The general statement can certainly be made, however, that there is no pervasive trend in the expected direction of decrease in dissemination correlating with decrease in frequency. Future work will have to address more fully whether meaningful statistical generalizations can be made about dissemination dynamics of receding words. However, as we explain in *Research Objectives*, we believe that item-specific explanations beyond general statistical tendencies are necessary for a sociolinguistically accountable discussion.

Contextualization Against the Sociolinguistic Record

From a sociolinguistic perspective, we do not necessarily consider the above findings alarming. The relationship between dissemination and frequency change is a statistical tendency that has been shown to hold as a generalization over large numbers of emerging words. Yet the dynamics of individual words are governed by more than the global statistical properties identified in Altmann et al. (2011), Altmann et al. (2013), or Stewart and Eisenstein (2018). In the case of *whom*, we have access to item-specific explanatory factors, such as the erosion of the English case system and the prolific metalinguistic discourse surrounding the word. As such, we can relate the change observable in the dissemination measures to this information in order to more fully understand the pathway of *whom* over the past 200 years. In this perspective, the dissemination measures are recontextualized as heuristic tools rather than variables used to test generalized hypotheses.

Returning to the extant literature on *whom*, a sketch of three developmental stages over the past three centuries can be drawn that is both in line with findings from previous research and able to account for dissemination developments as presented in our results. This sketch sees *whom* develop from 1) a carrier of grammatical information that is categorically required in a well-defined number of linguistic contexts to 2) a sociolinguistic variable that increasingly acquires stylistic over grammatical constraints and, finally, 3) a vestigial element which hardly shows productive variation in usage but retains salience thanks to active metalinguistic debate.

According to Aarts (1994), grammarians of the 18th and early 19th century treated variation between *who* and *whom* as a clear-cut case of complementary distribution: The former was reserved as the subject relativizer in RCs with human antecedent, whereas the latter was required both for direct objects and for complements of prepositions. That prescriptivist authors felt the need to formulate such a rule hints at the fact that there was some variation even in the 18th century, but at the same time the precept "became one of the most popular prescriptive rules in English grammar" (Aarts, 1994: 73). Its continued sway until well into the 20th century can be inferred from Sapir (1921: 166–174), who expresses unease at diverging from the normative pattern while at the same time recognizing the clear drift of English grammar away from *whom*.

During this first idealized stage, while grammatical context provides an unambiguous criterion for the choice between *who* and *whom*, the latter enjoys relative safety from the factors conspiring to ultimately lead to its demise: the overwhelming loss of nominal case inflection and the encroachment of other relativizers into its territory. However, the categorical, purely grammatical rule gradually morphed into a more context-sensitive one. Aspects of style were taken into account alongside, and increasingly: above, questions of case agreement. When and how precisely this change occurred has not yet been fully documented; Aarts (1994: 73) cites examples from 1985 onwards, leaving a gap of roughly 150 years to the most recent example of the former, rigid grammatical rule (Cobbett's *A Grammar of the English Language* from 1818).

Irrespective of the precise chronology, which can be assumed to have taken a gradual development at any rate, the relaxing of strict grammatical constraints made *whom* available as an indexically marked choice. A look at our data, and specifically: the development of register dissemination in **Figures 4** and **5**, suggests that the relaxing of the rule must have been in operation in the early 19th century already. At this time, *whom* is associated with elaborate and verbose texts, as can be inferred from the positive coefficients for the dimension STRUCTURAL ELABORATION and the negative ones for INFORMATION DENSITY. In other words, stylistic constraints had already come to play an important role around 1830.

In the long run, these associations likely did not help *whom* to retain much of its vitality. Throughout the recent history of English, there have been pervasive trends towards more efficient packaging of information and structural simplicity (Leech et al., 2009). As such, the rapid frequency decrease *whom* experiences between around 1850 and 1950 appears plausible. What is more puzzling at first glance is the concomitant increase in both linguistic and register dissemination. As to the former, the weakening of strict grammatical conditioning offers an explanation. While generally becoming less frequent, the occurrence of *whom* can no longer be predicted entirely from its immediate syntactic context. This is the case for RCs that formerly would have allowed no alternative to *whom*, but start to increasingly occur with *who*. Yet, perhaps more important are constructions in which *whom* would not have been permissible previously, but where hypercorrect application of an increasingly intransparent grammatical rule leads to its occasional, erratic appearance. Such hypercorrect usage is attested, among others, in Sledd (1987) and Tabbert (1990).

The increasing register dissemination visible throughout the period we analyze (see **Figure 4**) is an early sign of the last stage in our schematic representation, the retreat of *whom* from the vernacular grammar of most native speakers. Combined with the continued frequency decrease, the even register dissemination by the latter half of the 20th century suggests that *whom* simply is hardly used anymore at all, regardless of particular stylistic properties of individual texts. We observe its use mainly in two kinds of context: first, a small set of grammatical constructions that offer no easy alternative to replace *whom* by *who*, such as (2), and second, metalinguistic instances where *whom* is the subject of discourse rather than simply part of the discourse itself. Examples like the *who–whom–whomst–whomst'd* meme cited above, or a recent book entitled *A World Without* Whom*: The Essential Guide to Language in the BuzzFeed Age* (Favilla, 2017), highlight this latter usage.

(2) The characters, **between whom** the distances are long and harsh <COHA_mag_1989_486754>

Our interpretation that, by the late 20th century, *whom* is no longer a productive element in the active competence of most native speakers of English is in line with Bauer's (1994: 77) observation that readers in the late 20th century will generally recognize correct use of *whom* in a far wider range of contexts than they actually use it. It also finds confirmation in Mair (2006: 143) assessment that "*whom* is moribund as an element of the core grammar of English, but is very much alive as a style marker whose correct use is acquired in the educational system." Lasnik and Sobin's (2000) proposal to treat *whom* as a "grammatical virus" extraneous to vernacular grammar also hits a similar line.

This sketch of the historical development *whom* has undergone over the past 200 years, then, is able to accommodate the linguistic and register dissemination developments identified in our results. We summarize these relationships in **Table 3**. The role of social and topic dissemination remains less clear, partly because neither measure appears to correlate with frequency developments of *whom* in any meaningful way. It is possible that discourse topic simply has little bearing on the choice of relativizer, but we would expect social dissemination to yield clearer results, at least at times during which *whom* starts to acquire stylistic meanings. The most likely explanation for the absence of any clearer findings, we believe, lies in the nature of the data. With relatively few texts per year, especially in the earlier half of COHA, estimates of social dissemination suffer from considerable noise. The spread of individual points in **Figure 2** is a sign of this problem. We would expect a larger corpus database to offer clearer results.

Conclusion and Outlook

We have tested the association between frequency developments and changes in a range of word dissemination measures in the

TABLE 3 | Three developmental stages of whom.

Period	Status of *whom*	Sociolinguistic effects	Dissemination developments
Prior to 19th century	Regular grammatical conditioning	Stability due to categorical rules for relativizer choice	Not covered by our data; hypothesized stability of dissemination
19th and early 20th century	Predominantly stylistic conditioning	Variability between *who* and *whom; who* encroaches upon traditional *whom* contexts; hypercorrect use of *whom* due to intransparent grammatical rule	Low register dissemination indicating stylistic specificity; increase in linguistic dissemination as a consequence of weakening grammatical conditioning
1950s onward	Retreat from active use	Avoidance of *whom*; active metalinguistic discussion; discrepancy between awareness and use	Even dissemination due to overall low frequencies in all contexts

case of one receding word, *whom,* on the basis of historical corpus data comprising 180 years of written American English. In addition to the established metrics social dissemination (Altmann et al., 2011; Altmann et al., 2013) and linguistic dissemination (Stewart and Eisenstein, 2018), we have introduced two novel measures to quantify the dissemination of a word across registers and topics. The significant positive correlation between frequency and dissemination attested in the literature on emerging words was not found to hold for receding features.

Of the four measures we have considered, only linguistic and register dissemination showed meaningful changes between 1830 and 2009. These proved difficult to interpret in terms of general statistical tendencies, but became plausible once the specific sociolinguistic history of *whom* was considered. We proposed a trajectory of development, according to which *whom* changed from a regular grammatical to a predominantly stylistic marker and, in the latter half of the 20th century, to an unproductive element whose salience far supersedes its actual use.

Following Squires (2014), we submit that the relationship between dissemination and word vitality is best not conceived as static, but may assume different shapes at different stages of development. A productive goal for future research will be to reconcile this flexibility with an analytical perspective that goes beyond isolated, contextualized words. We are currently exploring unsupervised learning methods to find natural groups of words, based on their frequency and dissemination profiles across time. For instance, among the 50 most rapidly receding surface forms in COHA, we find some linearly declining social dissemination developments (e.g. for *nor* and *shall*), some random fluctuations around relatively constant values as with *whom* (e.g. also for *borne* and *circumstance*) as well as more complex, curvilinear trajectories (e.g. for *till* or *subject*). Based on a matrix of frequency information as well as (social, linguistic, register, and topic) dissemination at different intervals for individual words, we are working towards clustering words into groups that show similar developments over time.

The question remains why social and topic dissemination appear stable throughout the dynamic development of *whom* sketched above. D^T may simply not play an important role in general. The effect of discourse topic on linguistic variation is currently not well understood, as sociolinguists have often preferred to focus their analyses on different ways of saying the same thing (Labov, 1972: 271) rather than differences in what people talk about. More active consideration of discourse topic as a predictor of variation in sociolinguistics in general would be necessary to better interpret our findings in relation to topic dissemination.

As for social dissemination, it is difficult to accept that no notable change occurs alongside the decline of *whom* between 1830 and 2009. We have argued above that the measure may yield unstable results if the amount of individual texts is not sufficiently large, as reflected in the wide spread of yearly D^S values in **Figure 2**. Unfortunately, this property makes the measure problematic for many sociolinguistic applications, for which often only relatively small corpora are available. By contrast linguistic dissemination is able to draw on information from every instance of a feature's use and our new metric of register dissemination uses fine-grained, scalar information at the level of individual texts. Consequently, we expect both these measures to be better suited for application to comparatively small data sets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.english-corpora.org/coha/ for online access to the corpus itself https://github.com/axboh/WHOM_Dissemination for all intermediate datasets created for the analyses presented in the paper.

AUTHOR CONTRIBUTIONS

LH and AB were responsible for developing the theoretical background on the relativizer *whom*. MB and AB collaborated in implementing the social dissemination measure for words in COHA and developing the general theoretical background to word dissemination. AB calculated linguistic dissemination for *whom* in COHA and developed the register and topic dissemination measures. The text of the manuscript was in large parts produced by AB, with contributions from MB and LH.

FUNDING

The authors gratefully acknowledge funding from the Volkswagen Foundation for the project "Language as a complex adaptive System: Insights from physical modelling."

REFERENCES

Aarts, F., and Aarts, B. (2002). "Relative *Whom*: 'A Mischief maker.'," in *Text Types And Corpora: Studies in Honor of Udo Fries*. Editors A. Fischer, G. Tottie, and H. M. Lehmann (Tübingen, Germany: Narr), 123–130.

Aarts, F. (1994). Relative *Who* and *Whom*: Prescriptive Rules and Linguistic Reality. *Am. Speech* 69 (1), 71. doi:10.2307/455950

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2011). Niche as a Determinant of Word Fate in Online Groups. *PLoS One* 6 (5), e19009. doi:10.1371/journal.pone.0019009

Altmann, E. G., Whichard, Z. L., and MotterMotter, A. E. (2013). Identifying Trends in Word Frequency Dynamics. *J. Stat. Phys.* 151 (1–2), 277–288. doi:10.1007/s10955-013-0699-7

Bauer, L. (1994). *Watching English Change: An Introduction To the Study of Linguistic Change in Standard Englishes in the Twentieth Century (Learning About Language)*. Harlow, United Kingdom: Longman.

Bhattacharyya, A. K. (1943). On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. *Bull. Calcutta Math. Soc.* 35, 99–109.

Biber, D., and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge, United Kingdom: Cambridge University Press.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge, NY: Cambridge University Press.

Biber, D. (2012). Register as a Predictor of Linguistic Variation. *Corp. Linguist. Linguist. Theor.* 8 (1), 9–37. doi:10.1515/cllt-2012-0002

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text With the Natural Language Toolkit*. Beijing, Cambridge, MA: O'Reilly Media.

Blei, D., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Bohmann, A. (2019). *Variation in English Worldwide: Registers and Global Varieties (Studies in English Language)*. Cambridge, United Kingdom: Cambridge University Press.

Davies, M. (2010). Corpus of Historical American English (COHA). Available at: https://www.english-corpora.org/coha/ (Accessed December 20, 2020).

Egghe, L. (2007). Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. *J. Am. Soc. Inf. Sci.* 58 (5), 702–709. doi:10.1002/asi.20524

Favilla, E. J. (2017). *A World without "Whom": The Essential Guide to Language in the BuzzFeed Age*. New York: Bloomsbury.

Garley, M., and Hockenmaier, J. (2012). "Beefmoves: Dissemination, Diversity, and Dynamics of English Borrowings in a German Hip Hop Forum," in Proceedings of the Association of Computational Linguistics, Jeju Island, Korea, July 2012 (Jeju Island, Korea: Association for Computational Linguistics), 135–139.

Grieve, J. (2018). "Natural Selection in the Modern English Lexicon," in Proceedings of the 12th International Conference on the Evolution of Language (Evolang12), Torun, Poland, April 17, 2018 (Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika).

Grieve, J., Nini, A., and Guo, D. (2017). Analyzing Lexical Emergence in Modern American English Online. *English Lang. Linguist.* 21 (1), 99–127. doi:10.1017/S1360674316000113

Guy, G. R., and Bayley., R. (1995). On the Choice of Relative Pronouns in English. *Am. Speech* 70 (2), 148–162. doi:10.2307/455813

Hinrichs, L., Szmrecsanyi, B., and Bohmann, A. (2015). *Which*-hunting and the Standard English Relative Clause. *Language* 91 (4), 806–836. doi:10.1353/lan.2015.0062

Kretzschmar, W. A. (2015). *Language and Complex Systems*. Cambridge, United Kingdom: Cambridge University Press.

Labov, W. (1994). *Principles of Linguistic Change (Vol. 1: Internal Factors) (Language in Society 20)*. Cambridge, MA: Blackwell.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Lasnik, H., and Sobin, N. (2000). The *Who/Whom* Puzzle: On the Preservation of an Archaic Feature. *Nat. Lang. Linguist. Theor.* 18 (2), 343–371. doi:10.1023/A:1006322600501

Leech, G. N., Hundt, M., Mair, C., and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study (Studies in English Language)*. Cambridge, NY: Cambridge University Press.

Levey, S. (2006). Visiting London Relatives. *English World-Wide* 27 (1), 45–70. doi:10.1075/eww.27.1.04lev

Mair, C. (2006). Twentieth Century English: History, Variation and Standardization *(Studies in English Language)*. Cambridge, United Kingdom: Cambridge University Press.

Python Software Foundation (2020). Python. Available at: http://www.python.org. (Accessed December 20, 2020).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project-org/ (Accessed December 20, 2020).

Řehůřek, R., and Sojka, P. (2010). "Software Framework for Topic Modelling with Large Corpora," in Proceedings of the LREC 2010 Workshop on New Challenges for NLP, Valetta, Malta, May 22, 2010, 46–50.

Revelle, W. (2020). psych: Procedures for Psychological, Psychometric, and Personality Research. Available at: https://CRAN.R-project.org/package=psych (Accessed December 20, 2020).

Röder, M., Both, A., and Hinneburg, A. (2015). "Exploring the Space of Topic Coherence Measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, February 5, 2015 (Shanghai: ACM), 399–408.

Sapir, E. (1921). *Language*. London, United Kingdom: Harvest.

Sledd, J. (1987). The Whoming Pigeon. *Am. Speech* 62 (4), 379–380. doi:10.2307/455417

Squires, L. (2014). From TV Personality to Fans and Beyond: Indexical Bleaching and the Diffusion of a media Innovation. *J. Linguist. Anthropol.* 24 (1), 42–62. doi:10.1111/jola.12036

Stewart, I., and Eisenstein, J. (2018). "Making "Fetch" Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, November 4, 2018 (Brussels, Belgium: Association for Computational Linguistics), 4360–4370.

Tabbert, R. (1990). Rare Whoming Pigeon Sighted in the grove of Academe. *Am. Speech* 65 (2), 164–165. doi:10.2307/455536

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. Washington, DC: American Psychological Association.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/frai.2021.642783

Perception in Black and White: Effects of Intonational Variables and Filtering Conditions on Sociolinguistic Judgments With Implications for ASR

Nicole R. Holliday *

University of Pennsylvania, Philadelphia, PA, United States

This study tests the effects of intonational contours and filtering conditions on listener judgments of ethnicity to arrive at a more comprehensive understanding on how prosody influences these judgments, with implications for austomatic speech recognition systems as well as speech synthesis. In a perceptual experiment, 40 American English listeners heard phrase-long clips which were controlled for pitch accent type and focus marking. Each clip contained either two H* (high) or two L+H* (low high) pitch accents and a L-L% (falling) boundary tone, and had also previously been labelled for broad or narrow focus. Listeners rated clips in two tasks, one with unmodified stimuli and one with stimuli lowpass filtered at 400 Hz, and were asked to judge whether the speaker was “Black” or “White”. In the filtered condition, tokens with the L+H* pitch accent were more likely to be rated as “Black”, with an interaction such that broad focus enhanced this pattern, supporting earlier findings that listeners may perceive African American Language as having more variation in possible pitch accent meanings. In the unfiltered condition, tokens with the L+H* pitch accent were less likely to be rated as Black, with no effect of focus, likely due to the fact that listeners relied more heavily on available segmental information in this condition. These results enhance our understanding of cues listeners rely on in making social judgments about speakers, especially in ethnic identification and linguistic profiling, by highlighting perceptual differences due to listening environment as well as predicted meaning of specific intonational contours. They also contribute to our understanding of the role of how human listeners interpret meaning within a holistic context, which has implications for the construction of computational systems designed to replicate the properties of natural language. In particular, they have important applicability to speech synthesis and speech recognition programs, which are often limited in their capacities due to the fact that they do not make such holistic sociolinguistic considerations of the meanings of input or output speech.

Keywords: intonation, sociophonetics, African American English, ethnic identification, automatic speech recognition, speech synthesis

OPEN ACCESS

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Kathryn H Franich,
University of Delaware, United States
Kevin Tang,
University of Florida, United States
Delin Deng,
University of Florida, United States, in collaboration with reviewer KT
Rachael Tatman,
Rasa Technology, United States

****Correspondence:***
Nicole R. Holliday
nholl@sas.upenn.edu

Received: *16 December 2020*
Accepted: *30 June 2021*
Published: *15 July 2021*

Citation:
Holliday NR (2021) Perception in Black and White: Effects of Intonational Variables and Filtering Conditions on Sociolinguistic Judgments With Implications for ASR.
Front. Artif. Intell. 4:642783.
doi: 10.3389/frai.2021.642783

INTRODUCTION

The questions of whether and how listeners can distinguish Black American and White American voices have been a popular topic in phonetic and sociolinguistic studies over the past 50 years, with implications for both the linguistic understanding of perception as well as issues of social inequality (see for review Thomas and Reaser 2004; Thomas et al., 2010). In general, these studies have found that listeners are fairly adept at distinguishing Black American and White American voices, though the literature has not yet completely established which acoustic parameters may influence listeners' judgments. In particular, though scholars have posited distinctive patterns of intonation, prosody, and voice quality associated with varieties of African American Language[1], the specific acoustic characteristics of these varieties are still not well-described. This is a serious lacuna, because as a result of their perceptual salience, intonational features are especially important in the analysis of linguistic profiling, or what noted linguist John Baugh has recently called "Speaking While Black", the phenomenon by which African Americans experience discrimination, sight-unseen, because their speech may act as an indicator of their race (2015).

As scholars such as Baugh (2000, 2003, 2015) and Thomas and Reaser (2004) have noted, understanding the ways in which listeners make ethnicity judgments is crucial for working against discrimination and linguistic profiling. In his 2015 chapter in the Oxford Handbook of African American Language (OHAAL), Baugh provides ample evidence of this type of discrimination in the courtroom, in housing, and in the workplace. Indeed, though we know THAT listeners make these judgments, the question of precisely HOW still escapes sociolinguists. This understanding is of vital importance due to the fact that until we know how linguistic profiling occurs, we will not be able to provide professionals across industries as well as the public with strategies to recognize and combat this type of discrimination. As Baugh observes, "it is important that those who speak non-dominant dialects or non-dominant languages are aware of their linguistic circumstances, but also the constraints they may face from those who are fluent speakers of surrounding dominant languages and dialects" (768). As a result, linguists have a powerful motivation to better understand the scientific mechanisms that underlie their social judgments about language.

Traditionally, much of the literature that examines the way in which varieties of African American Language are stigmatized has focused on phonological and morphosyntactic differences between AAL varieties and Mainstream U.S. English (MUSE) (Spears 1988; Thomas 2015). However, some research has indicated that speakers can be reliably identified as Black by listeners, even in the absence of non-standard grammatical features (Purnell et al., 1999; Thomas and Reaser 2004; Holliday and Jaggers 2015). In these contexts, even Black speakers who do not use stereotyped variables associated with AAL may still be subject to linguistic profiling and discrimination due to their use of intonational and prosodic features that index Blackness in the minds of listeners. This substantial gap related to the study of intonational features in the literature represents a serious challenge for both linguists and lay people alike, especially given that suprasegmental features are among the elements of speech that are most salient for listeners, even if they are not consciously aware of this fact (Thomas 2015).

Beyond issues related to linguistic profiling by humans in real-life situations, the lack of research on ethnolinguistic variation at the level of intonation also represents a challenge for scholars interested in how to employ listener judgment and production data for computational applications. A number of recent studies have begun to show the limitations of assuming ethnolinguistic homogeneity in language recognition and synthesis programs, and have advocated for addressing the role of several types of bias in NLP applications (Blodgett et al., 2020; Shah et al., 2020). Though the majority of this research has focused on large-scale corpus data such as tweets, the issues may be even more pressing for the analysis of spoken data. Tatman and Kasten (2017) tested effects of talker gender and race on automatic speech recognition in two (ASR) systems: Bing Speech and Youtube automatic captions, and found a significantly higher word error rate for African American talkers than for White ones. In this way, Black American speakers experience linguistic discrimination not just by humans, but also by systems designed to process human language, as systems that do not consider ethnolinguistic variation are more likely to fail them. However, this outcome is not inevitable: Lehr et al. (2014) found that specifically training a discriminative pronunciation model on AAVE data improved the model's accuracy by 2.1%, showing that with proper data and training, systems can begin to accommodate complex ethnolinguistic variation.

In particular, as speech recognition and speech synthesis systems become more integrated, having a better understanding of the criteria that listeners use when making social judgements under different types of listening conditions thus is important for improving the quality of models. Incorporating information about ethnolinguistic variation may be especially vital for researchers working in the area of speech synthesis. Individuals with medical conditions that impact their speech frequently rely on systems like Speech Generating Devices (SGD), and patients show strong preferences for systems that generate voices that align with their social identities (Crabtree et al., 1990; Creer et al., 2013). To date, few existing systems take into account factors such as ethnolinguistic variation, but having a naturalistic voice output system has been shown to improve the quality of life for patients with these types of conditions (Creer et al., 2013).

The current study begins to address these issues by examining two suprasegmental parameters that have been observed to be involved in such ethnolinguistic variation; pitch accent type and focus marking, as potential loci of information that listeners may use to make ethnicity judgments. By focusing on these two variables, which have been observed to differ between Black and White speakers in production studies (Author 2016; McLarty 2018), linguists may be able to start to pinpoint the intonational variables that influence listener ethnicity judgments.

[1] I used AAL here as a cover term for several varieties of English spoken in Black American communities, following Lanehart (2015).

This study also builds on earlier works on production to investigate the relationship between variables that speakers use to perform certain types of racial identity as well as listener judgments of those same variables. It aims to corroborate the observations in earlier studies that have found listeners accurate at judging ethnicity, but also to carefully control the intonational phenomena in the stimuli to investigate the role of those variables in these types of judgments. It also challenges some of the assumptions made in earlier ethnicity judgment studies by showing that intonational contours may be judged differently by listeners when they are exposed to filtered vs. unmodified speech.

The majority of previous studies on ethnic identification exposed listeners to unaltered sound clips that were made in a laboratory setting, and also asked listeners to make judgments in a laboratory setting under ideal listening conditions (cf Thomas and Reaser 2004). One limitation of this methodology is that laboratory listening conditions may differ from the everyday type of listening environments where linguistic profiling happens. The current study's findings provide further evidence that linguists should consider the potential effects of listening environment since the results indicate that listeners may pattern in opposite ways with respect to ethnicity judgments, depending upon the intonational contours of the stimuli as well as how the stimuli may be filtered. This difference in judgments based on intonation and filtering may be even more important to understand for use in computational systems, given that the corpora employed frequently employ data that has been recorded under imperfect acoustic conditions that may have filtering effects, such as YouTube Videos (Tatman and Kasten 2017). Furthermore, for speech synthesis applications, users necessarily interact with listeners in imperfect acoustic settings, indicating that the creation of more naturalistic synthesized voices must consider the way listeners evaluate speakers in a variety of conditions.

PERCEPTION, RECOGNITION, AND PRODUCTION OF ETHNOLINGUISTIC VARIATION

Perception Studies

As observed by Thomas (2015) and Author (2016), the overall lack of information about the role of suprasegmental features, such as intonation and prosody, in the speech of Black Americans presents an important challenge for researchers. Despite the evidence that prosodic information is highly salient for listeners when making judgments about speaker ethnicity, we still have very little information about how different acoustic parameters may affect these assessments (Purnell et al., 1999). This also presents particular difficulty for inclusive ASR systems; if we do not understand the parameters that listeners use to distinguish voices, we cannot properly evaluate systems that should be able to respond to the variation inherent in large communities of users. In particular, recent research on bias in NLP models reveals tendencies to exclude or stereotype language employed by Black users, leading to communities being not only underserved but also harmed by such systems (Blodgett et al., 2020). Sociolinguistic research on such bias as well as how listeners interpret social properties of voices may be able to help researchers in ASR and synthesis begin to address these inequalities.

In their 2004 study and review of the literature on ethnic identification, Thomas and Reaser discuss 30 studies from across the U.S. that have generally supported the finding that American English speakers are adept at identifying the race of a speaker, even based on hearing a very short sound clip, indicating the role of features beyond the segmental level. Although AAL suprasegmental features (and intonational features in non-standard English varieties overall) have received less attention than other types of phonological or morphosyntactic features, a few scholars have addressed the role of these features in ethnicity judgments. In this vein, there are a number of studies that have approached this type of variation from a production perspective, and others that have addressed it from a perception perspective. In general, the production studies have been more likely to employ methodologies designed specifically to test the variables involved in ethnolect variation, although this can also be observed in some of the perception studies that have conducted posthoc analyses of the variables involved. The section that follows will begin by describing the findings of earlier perception studies, and will then discuss the findings of relevant production studies on listener judgments of ethnicity.

In one of the earliest ethnic identification studies in the U.S., Bryden (1968) analyzed Black and White speakers in Charlottesville, Virginia, and found that ethnicity was correctly identified for 74% of speakers in unfiltered stimuli, and correctly identified for bandpass filtered stimuli approximately 68% of the time. In this study, listeners heard unfiltered recordings of racially matched speakers and then heard the same tapes again that had undergone bandwidth compression using spectral filtration. The stimuli were taken from 35 White and 35 Black children reading the United States' Pledge of Allegiance. In the filtered listening condition, stimuli were band pass filtered below 1,250 and above 1,750 Hz. Bryden motivates this level of band pass filtering by claiming that a filter between 1,250 and 1,750 Hz is the maximum filtering condition that can be employed without loss of intelligibility. The listeners heard 20 filtered and 20 unfiltered clips each and the listening population included 40 listeners, 20 of whom were Black and 20 of whom were White, and 8 of whom had some previous training in the field of communication sciences and disorders. Bryden's primary finding was that listeners' ability to make accurate ethnicity judgments is somewhat degraded in filtered conditions, but that listeners still performed better than chance even in these bandpass filtered conditions, showing the durability of listener judgments even with degraded stimuli.

Building on this work, Koutstaal and Jackson (1971) examined ratings of the voices of 10 male speakers in Ohio, and found that speakers were over 80% accurate in their identifications, though they were somewhat more accurate with White speakers than Black speakers. 26 listeners heard "five negro colloquialisms"[2]

[2] To a modern reader, these may be colorfully dated phrases. "None of that off the wall stuff", "What's happ'nin man", "Man I don't play the dozens," "Let's go grease," and "She ain't nothing but a stone fox".

that had been read by 5 Black and 5 White Ohio speakers, and were instructed to simply indicate whether the speaker was Black or White. Results indicated that the White speakers were almost categorically identified as White (4 were identified at 100% accuracy and one at 92% accuracy). However, the authors observed substantially more variation for the Black speakers, with two of the Black speakers identified at 100% accuracy and the others identified at 85, 81, and 69%, respectively. Koutstaal and Jackson (1971) also conducted a posthoc analysis of the samples using spectrograms, specifically examining syllable times, overall speaking time and F0 for each clip. They identified the presence of different contours for Black and White speakers, but concluded that contour shape was not predictive of speaker identification[3]. They also found no consistent differences for syllable duration or any of the other suprasegmental variables that they examined. Ultimately, the authors speculate that listeners may use segmental information rather than intonational information in their judgments, but they do not systematically evaluate these differences.

Lass et al. (1980) conducted a study in which listeners heard sentences from 10 male and 10 female speakers that had been read in an unfiltered condition, low pass filtered at 255 Hz, and high pass filtered at 255 Hz. These filtering conditions were used to attempt to focus on which variables listeners may attune to in their judgments. A low pass filter at 255 has the effect of eliminating vowel formant while retaining fundamental frequency, while a high pass filter at 255 Hz has the effect of focusing listener attention on vowel formants, though the signal still retains traces of F0 information (Thomas 2011:76). Lass et al. (1980) found that listeners correctly identified speaker ethnicity 72% of the time in the unfiltered condition and that identification rates were lower but still reliable for the other conditions, with accuracy rates of 69% for the high pass filtered condition and 60% in the low pass filtered condition. The authors concluded, based on these results, that formants are generally more important than F0 measures for ethnic identification, given the higher accuracy rate with the high pass filter.

In addition to their review of the earlier literature, Thomas and Reaser (2004) also conducted an experiment where they examined ethnic identification with Black and White American speakers from Hyde County, North Carolina, as well as inland regions of North Carolina. In their experiment, 117 listeners rated three different types of clips. In one condition, the clips were unmodified. In the second condition, they were monotonized using KayAnalysis Synthesis Laboratory with F_0 set at 120 Hz for male speakers and 200 Hz for female speakers, in order to eliminate F_0-dependent variation. In the third condition, the stimuli were low pass filtered at 330 Hz, in order to preserve intonational information while removing nearly all vowel quality cues. Their results indicated a high level of accuracy for the monotonal treatment among all listeners, and a rate of accuracy close to chance for the low pass filtered conditions. Thomas and Reaser (2004) do note that the filtered stimuli containing prominent subject pronouns were more readily identified than those without such pronouns, indicating that listeners may be relying on at least some intonational information in making their judgments, though they were not specific about what types of intonational contours occurred in these contexts.

Foreman (2000) found that listeners were over 80% accurate in ethnic identification and that those listeners with greater exposure to both White and Black voices were the most accurate. In this study, she tested 20 Black listeners and 19 White listeners on recordings of a script made by 6 Black and 4 White speakers. The stimuli consisted of 54 sentences with “distinctive intonational patterns”, though Foreman is not explicit about what these intonational patterns were. The stimuli were low pass filtered at 900 Hz to partially obscure segmental and voice quality cues in order to specifically test the role of intonation. It is important to note that in contrast with the study conducted by Lass et al., this filter setting at 900 Hz still allows for some formant information as well as a higher level of intelligibility of the signal. Foreman notes that the sentences with “ethnically diagnostic intonation patterns” were most easily identified, though she does not state exactly which contours she tested. Despite the fact that Foreman is not specific about which “dialect specific” intonational contours she employed in the stimuli, she does claim that stimuli with “distinctively” Black intonation are more likely to be correctly identified has having been uttered by a Black speaker, providing evidence for the importance of intonational variation in these judgments. Supporting the patterns also observed by Buck (1968) and Koutstaal and Jackson (1971), Foreman also found that listeners were less accurate in identifying Black speakers than White speakers, a finding that she attributes to the fact that Black speakers may not always employ stereotypical AAL features in every utterance.

Foreman (2000) findings are especially important in light of how the results of these ethnic identification tasks may be important for computational linguistic applications. Foreman posits an expectation that listeners are waiting to hear stereotypical AAL features, and that when they do not, they have lower levels of accuracy in the ethnic identification task. Given that Lehr et al. (2014) were able to improve the accuracy of an ASR model by training it on phonological and morphosyntactic features of AAL, it may be reasonable to hypothesize that such training on intonational features may provide even greater improvements to such models. Since ASR systems can be trained to examine features at all linguistic levels, not just those with stereotypical salience, understanding the role of prosodic variation may allow such systems to improve on listener-ratings, if the systems can be trained to avoid the pitfalls of stereotypes that listeners may experience. In this way, examining the performance of ASR systems at different levels of filtering may also help us better isolate which variables may be more or less salient for human listeners in similar tasks.

Perception and ASR Systems

With respect to how such ethnicity judgments may affect the performance of speech recognition and synthesis systems, little work has specifically explored how such systems may evaluate ethnic differences between inputs. In fact, not only is there a

[3] It is unclear whether any of their contours correspond directly to the ToBI H* or L+H* employed in the current study.

dearth of literature examining how ASR systems may incorporate sociolinguistic information at any level, there is also very little that directly compares how humans and systems incorporate different acoustic information in ways that may be similar or different from each other. This is a serious problem for researchers across fields, because as Blodgett et al. (2020) observe "work must be grounded in the relevant literature outside of NLP that examines the relationships between language and social hierarchies; without this grounding, researchers and practitioners risk measuring or mitigating only what is convenient to measure or mitigate, rather than what is most normatively concerning" (6). Despite this limitation related to disconnects between linguists and computational researchers and lack of research comparing human and machine performance, some work has begun to address issues related to both managing variable inputs and accounting for noise in ASR systems. Unfortunately, many modern systems rely on proprietary deep learning algorithms for speech recognition and generation, so the properties of these systems are not necessarily transparent. Earlier foundational research, however, has discussed the mechanisms that underlie some of these processes, which will allow us to discuss how systems have addressed variable acoustic inputs.

Linguists and other researchers have long observed the necessity for naturalistic prosody in computational linguistic applications. In their 2010 paper, Vicsi and Szaszák, 2010 address acoustic processing and modelling of the suprasegmental speech properties and find that the addition of prosodic information, in this case F0 and energy, significantly improves word recognition and boundary detection in models for both Hungarian and Finnish. Their system begins with Hidden Markov Model units that they then train and connect to a broader language model. They subsequently use the HMM framework to model prosody and conduct syntactic and/or semantic level processing of the input speech and then used HMMs to model each clause's prosodic contour. They claim that the addition of prosodic contour modeling increased accuracy; for Hungarian data, word recognition improved by 3.82% with the addition of prosodic information.

In an early work on prosody modeling for ASR, Shriberg and Stolcke, (2004) examine a number of different strategies for incorporating prosodic information into their models. One of their main arguments relates to the fact that computational systems need not necessarily process and manage linguistic input the way that human coders do. In fact, they argue that computational systems should be modeled "directly in a statistical classifier—without the use of intermediate abstract phonological categories, such as pitch accent or boundary tone labels. This bypasses the need to hand-annotate such labels for training purposes, avoids problems of annotation reliability, and allows the model to choose the level of granularity of the representation that is best suited for the task" (2). Unfortunately, this creates a significant difference from how linguists interested in human speech model prosodic information, making the two approaches difficult to compare directly. However, these authors do provide important information about the criteria that many models are based on, noting that their method is based on contour classification on both syntactic and semantic models. In particular, they note that the most successful models that they observe "extract features from a forced alignment of the transcripts (usually with phone-level alignment information), which can be based on either true words, or on (errorful) speech recognition output…This yields a rich inventory of "raw" features reflecting F0, pause and segment durations, and energy (2). Though these models rely on statistical classifiers as opposed to the phonological categories used by non-computational researchers, the features that their model incorporates overlap significantly with the features that human coders use to do prosodic labelling. To date, I have found no research that directly compares human coders and statistical models for prosody, but Shreiber and Stolcke's findings provide support for the claim that ASR models may be trained to use the same type of phonetic criteria for prosodic labelling that human coders use (Cole and Shattuck-Hufnagel 2016). As a result, a better understanding of the phonetic cues that may differ between ethnolinguistic communities has the potential to enhance the accuracy of ASR systems as well.

As the current study is interested in how humans make ethnicity judgments under different listening conditions and how this may compare with computational systems, how ASR systems perform under noisy conditions is another important point of consideration. Li et al. (2014) provide a comprehensive overview of the literature on noise-robust ASR as well as a useful comparison that clearly articulates advantages and disadvantages of various popular models. While a full discussion of the five types of models they compare is beyond the scope of this review, they do provide some important points that are especially relevant to the current study. In particular, they compare systems that employ five different types of attributes: "feature vs. model domain processing, explicit vs. implicit distortion modeling, use of prior knowledge about distortion or otherwise, deterministic vs. uncertain processing, and joint vs. disjoint training" (768). Though many modern systems rely on neural-network based methods or CD-DNN-HMM, a number of the older methods continue to provide the basis for their assumptions; this is particularly the case for explicit distortion modeling. Li et al. argue that "noise, channel, and speaker factors may already be well normalized by the complex nonlinear transform inside the DNN. However, this does not mean that the noise-robustness technologies are not necessary when used together with CD-DNN-HMM" (771). This indicates that the authors believe that ASR can be improved when models are given explicit training on the speech context, including information about the speaker, which may also include the type of sociolinguistic information available to human listeners. Having examined both how humans use acoustic information to make judgments and how ASR systems use such information in their models, we now turn to the question of how the same acoustic variables have been examined in studies that focus on how humans produce speech.

Production Studies and the Tone and Break Index System

Though production studies on ethnicity and suprasegmental variables have also been somewhat rare, several have focused on observing systematic differences between Black and White speakers in a variety of speech settings. Unfortunately, few have tested whether these differences in production are truly salient for listeners, which is an especially important consideration for speech synthesis.

In terms of the specific intonational features that may differ between MUSE and AAL patterns, there have been only a few studies that have examined this question in a modern framework. Starting in the 1980's, intonational phonologists and phoneticians began to employ the modern Tone and Break Index transcription system for General American English (Pierrehumbert 1980; Beckman and Pierrehumbert 1986, cf ; Beckman et al., 2005). This system consists of an inventory of pitch contours (tones) and phrase boundaries (breaks) and is in widespread use in the modern literature on intonational phonology (ibid, Thomas et al., 2010; Thomas 2015, inter alia). It is especially useful for systematically examining variation and providing a consistent framework for labeling intonational contours and phrase boundaries, and so the majority of intonation studies published in the U.S. in the last 30 years have employed the ToBI annotation conventions. The full guidelines for TobI labelling can be found in Beckman and Ayers (1997), though for the current study and those discussed below, the primary points of interest in the ToBI system are pitch accents and boundary tones, which consist of a number of combinations of high (H) and low(L) tones, which can generally be seen in the shape of the F0 contour on a spectrogram. In English, pitch accents can only occur on stressed syllables, and they are the main cue to prominence. They are typically realized with a combination of some type of F0 movement as well as other cues such as longer duration and higher intensity. The pitch accents of interest for the current study will be discussed in greater detail in the methodology section, though understanding the basics of the framework is necessary for interpreting the findings of Jun and Forman, (1996), McLarty (2018), and Author (2016) which are discussed below.

Jun and Forman (1996) provide the first formal analysis of AAL intonation based on this autosegmental metrical model and the ToBI system (Pierrehumbert 1980). Jun and Forman (1996) recorded 7 same-race dyads (5 Black and 2 White), enacting the same scripted dialogue. They found that in general, the Black speakers (who were all speakers of AAE), employed wider pitch range and higher pitch at phrase boundaries than the White MUSE speakers. Specifically, they were interested in the patterning of Yes-No questions, and found that AAE speakers appeared to have a different pattern than the MUSE speakers, such that the AAE speakers were more likely to use a low tone followed by a high flat tone (L* H-L% in ToBI) while MUSE speakers use a low tone followed by a rising high boundary (L* H-H% in ToBI) though the differences between the two speaker groups were less consistent for declaratives and Wh-questions. This study represented an important first step for systematically analyzing the differences between MUSE and AAL using modern intonational techniques, though its design and focus on phrase boundaries and Yes-No questions limits its applicability for testing listener judgments of ethnicity using declaratives and naturalistic speech.

McLarty (2018) also attempted to quantify the specific differences between Black and White speakers with respect to intonational variables, using the ToBI framework. McLarty studied phenomena related to differences in pitch accent types, using the ex-slave recordings previously employed in linguistic research by Bailey et al., (1991) as well as contemporary speakers from Raleigh, North Carolina, McLarty found that African Americans in both the ex-slave and modern recordings used a greater incidence of the L+H* (low target followed by high target with prominence, in the same syllable) pitch accent as compared to the H* (high) pitch accent, when compared to the MUSE speakers in his study. McLarty argues that this may provide further evidence for a generally different pattern of use of intonational contours between MUSE and AAL speakers, though he also did not test the salience of these observed differences for listeners.

In my earlier study, Author (2016), I examined casual speech data from young men with one Black parent and one White parent, who I refer to as BWIs (Black/White individuals)[4]. Using sociolinguistic interview data and self-reported identity markers for participants, as well as a modification of the multiracial identity model proposed by Rockquemore et al. (2008), participants were examined for self-reported identity type as multiracial and/or Black. Participants were recorded in casual peer dyad conversations with friends, and the analysis of their intonational patterns was taken from these recordings. In this study, I found a general pattern such that the participants who identified more as Black, as opposed to multiracial or mixed, were more likely to use a greater quantity of L+H* accents than H* accents. This pattern parallels the findings of McLarty (2018), who found that AAL speakers were more likely to use more L+H*s than MUSE speakers. An example of these accents from this data set, which were also used as stimuli in the current study, can be observed in the Praat (Boersma and Weenink, 2014) spectrograms **Figure 1** and **Figure 2**. In particular, observe the movement of the pitch tracker over the course of the spectrogram.

As we can observe from **Figures 1**, **2**, these intonational contours are differentiated primarily by their shape, with the H* contour simply having a high target, and the L+H* contour having low to high movement all within the prominent syllable.

In addition to the finding that speakers who identified more as Black used a greater quantity of L+H* pitch accent, Author (2016) also found that speakers were more likely to employ L+H* in phrases with narrow, as opposed to broad, focus marking, which is a pattern that would be expected for speakers of MUSE. However, the speakers in Author (2016) who identified most strongly as Black also employed L+H* in broad focus conditions, which is not predicted in MUSE. In English, narrow focus is often

[4] Black/White Individuals. Participants in this study self-identify with a variety of racial categories, but it is important to note that the speakers in this sample initially only responded affirmatively to the question "do you have one Black parent and one White parent?" in the recruitment phase. For this reason, I have chosen to discuss them only by their response indicating their parentage, which allows us to discuss their external societal classifications and ancestries without ignoring their individual and nuanced self-descriptions as "multiracial", "biracial", "Black", "White", "other" or any combination thereof.

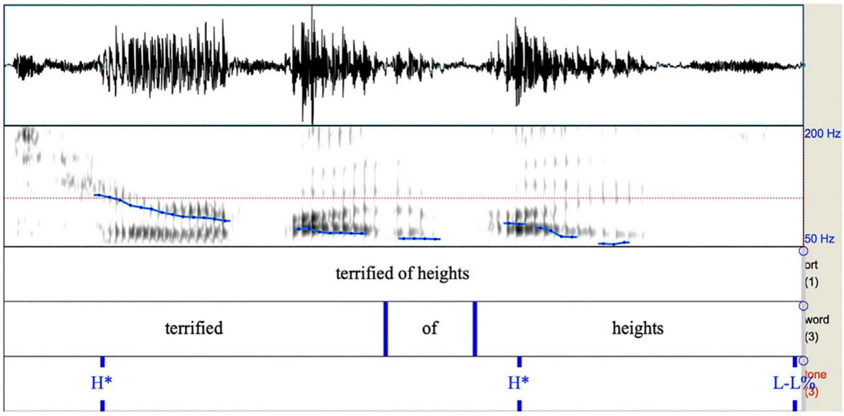


FIGURE 1 | Spectrogram of intonational phrase from Author (2016) containing two H* pitch accents followed by a L-L% boundary tone. Note the high F0 on the first syllable of terrified, and the rise on the first syllable of "heights".

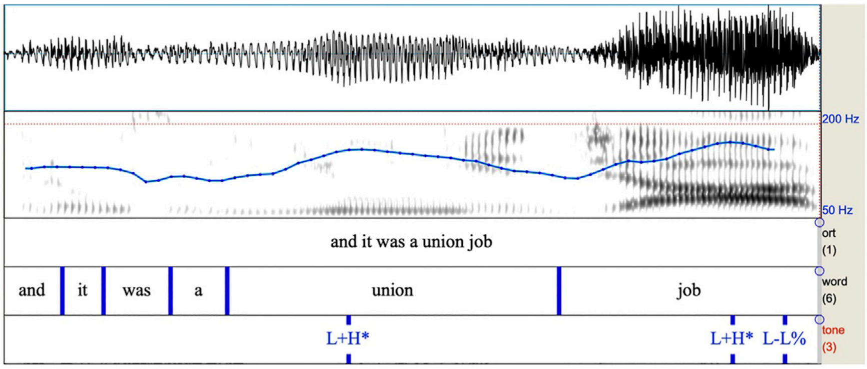


FIGURE 2 | Spectrogram of intonational phrase from Author (2016) containing two L+H* pitch accents followed by a L-L% boundary. Note the F0 fall and subsequent rise before the first syllable of "union". The same pattern occurs for the word "job".

thought of as contrastive, and it is characterized by the focused syllable (which must be the stressed syllable in the prominent word) being louder and longer than it would be if the same syllable appeared in broad focus. Though it is difficult to visualize on a spectrogram, focus can be reliably auditorily coded by listeners (Pierrehumbert and Hirschberg 1990). The primary use of narrow focus marking in MUSE appears to be to signal contrast, with speakers having been shown to employ narrow focus more often in situations where they need to indicate contrastive meaning, though this has not been through described or tested in AAL (ibid). Compare, for example, the following phrases, where capital letters are used to indicate narrow focus:

1. Jamal hugged Jim.
2. Jamal hugged LUKE.

Sentence 1 is a common type of phrase with broad focus. However, if we imagine a situation in which a listener hears Sentence 1, knows it to be incorrect, and would like to correct the speaker, the listener may utter Sentence 2, placing narrow focus (realized in part via longer duration and higher amplitude), to provide contrast with the incorrect assertion made in Sentence 1. Though Author (2016) found that L+H* was still likely to occur in such contexts, I also found that the speakers who identified most strongly as Black used L+H* in both broad and narrow focus conditions. While research on MUSE has found L+H* in narrow focus, L+H* in broad focus is not typically predicted, though it is possible that its use specifically in broad focus contexts may be characteristic of AAL.

To date, there is very little research on variation in strategies for focus marking in different varieties of English, though there is some literature indicating that it is theoretically possible since it is a site of variation in other languages. Frota (2002) observes that varieties of European Portuguese (EP) differ from varieties of Brazilian Portuguese in that they employ a specific contour (H*+L) to indicate narrow focus marking. Other languages, including Bengali and Italian, like EP, also use a special pitch accent to cue narrow focus, but languages such as English have not been observed to employ this strategy (ibid, Xu and Xu 2005). Jun and Forman (1996), also posit potential differences between AAL speakers and MUSE speakers with respect to focus marking, though they are not explicit about what these differences may be.

The findings in Author (2016) inspired a number of questions about the perceptual salience of these different pitch contours and focus marking strategies. Using data from a corpus built in that experiment, the current methodology is designed to test the hypothesis that listeners are more likely to rate tokens with an L+H* contour and/or narrow focus marking as having been uttered by a Black speaker. Understanding how these pitch accents and focus marking strategies are produced by speakers and perceived by listeners will help us arrive at a better understanding of the intonational phenomena that may trigger certain types of ethnicity judgments, as well as how these phenomena may be programable to assist computational systems in categorizing user data.

The current study is also unique because of its use of speakers with one Black parent and one White parent (BWIs) as opposed to White speakers and Black speakers. Previous studies on ethnic identification have focused on accuracy as a metric to identify which intonational factors may be salient to listeners, and these methods have generally ignored the rich variation that exists between Black speakers. This is especially limiting, due to the findings of Spears (1988) and Rahman, (2008) indicating that intonational factors may be the most important sites of variation for Black speakers who do not employ stereotyped features of AAL. In a way, this study follows in the path of Lambert et al. (1960) and others since who have employed a matched guise technique, with the primary difference being that the intonational phenomena are what distinguish the guises from one another. Everything else about the speakers' identities and voices is held constant, so in this way, we may arrive at a more precise understanding of the role of the intonational phenomena itself. This study pushes the field of ethnic identification forward both by using a previously unstudied speaker population, but also by pairing the stimuli that listeners hear specifically by intonational factors.

Following Bryden (1968), Koutstaal and Jackson (1971), Lass et al. (1980), and Thomas and Reaser (2004), this study also addresses the question of how listeners are affected by stimuli that have been altered using a specific type of low pass filter. These studies generally found that listeners were somewhat less accurate at identifying filtered stimuli than unfiltered stimuli, and the current study aims to test this with a new speaker population and as applied to clips that display specific intonational characteristics. As Thomas and Reaser (2004) note, the earlier studies, including their own, have the limitation of not necessarily corresponding to listening conditions in which real people make ethnicity judgments on an everyday basis. Though a lowpass filter may not replicate everyday listening conditions, examining differences between unmodified and filtered results may provide a more comprehensive understanding of how listening conditions affect judgments. The current study, alongside these earlier works, provides further motivation for the careful consideration of listening environment and noise when making claims about how listeners may evaluate speech across a variety of environments that may not resemble the listening conditions of lab speech. Furthermore, understanding both how speakers use these prosodic differences as well as how they are perceived by listeners will inform future research on naturalistic speech recognition and generation that functions more effectively for a wider variety of speakers and users.

METHODOLOGY

Stimuli

As a result of the fact that intonational variation between AAL and MUSE is still not well documented in the literature, and especially not using modern frameworks for understanding intonational variation, this experiment will act as a first pass at narrowing down the effects of both focus and pitch accent type on filtered and unfiltered speech types, as well as with different groups of listeners organized by race and gender. These results will contribute to our understanding of how certain aspects of ethnolinguistic variation are differentially perceived, and will assist in improving computational systems that necessarily must deal with variable production.

This study uses as stimuli six speakers from Author (2016) and asks listeners to rate the voices under different conditions. The corpus constructed in Author (2016) consists of recordings of young men with one Black parent and one White parent (BWIs), aged 18–32, who were recorded in Washington, D.C. or Eastern Virginia. All of the speakers self-reported that they are native speakers of both MUSE and AAL, though they were never explicitly instructed to speak in one variety or the other, as the original study was designed to explore the speakers" naturalistic range of intonational variation. In these recordings, the speakers are engaged in an "icebreaker"-style game with two different male-identified individuals (one White, one Black) that they identified as close friends. In this game, speakers were instructed to take turns asking each other questions (such as "What's the worst haircut you've ever had?" or "Describe your perfect afternoon") on cards for 20 min, though only the last 15 min of the recording were analyzed. Author (2016) did not find significant differences in patterns of intonation for these speakers by interlocutor, though I did observe significant differences conditioned by the speakers' attitudes about race and their own racial identities. The six speakers selected for the current study were the individuals who employed both the L+H* and H* tokens in large enough quantities to create the stimuli needed for the present experimental conditions.

Further information about the speakers' backgrounds, attitudes towards race, and more is available in Author (2016). A more thorough discussion of the speakers' characteristics is beyond the focus of the current study, though care was taken to select speakers who employed similar patterns of intonational variation to each other in the earlier study, as well as to control for potential speaker effects in the models of analysis. Though it may have been ideal to more tightly control for region, age, or interlocutor, given the limitations of intonational data and the fact that the target pitch accents do not necessarily appear with a high level of frequency for each speaker in each interlocutor condition, it was impossible to do so in the current study. Given the fact that Author (2016) did not find regional, age, or interlocutor differences with respect to the intonational variables of interest in this corpus, it is unclear whether further controlling for these variables in the creation of the stimuli for the current study would yield stimuli significantly different than those employed, since the study is designed primarily to test the perception of different intonational variants. However, replication experiments that employ stimuli from speakers that are more tightly controlled along these dimensions may be a useful avenue for future work.

Intonational phrases from the corpus created by Author (2016) were annotated using the aforementioned ToBI conventions for Mainstream English, in order to obtain phrases with the pitch accents types of interest, H* and L+H* (Beckman et al., 2005). From this corpus, the six speakers who all had phrases of the type of interest were selected. Listeners heard eight intonational phrase-long clips from each speaker in two experimental conditions (one low passed filtered and one unmodified), and the phrases were presented in a randomized order which was different for each listener.

For purposes of this experiment, an intonational phrase was classified as any phrase containing at least one pitch accent and a boundary tone, following the ToBI conventions of Beckman and Ayers (1997). Phrases were selected to be of comparable length (mean syllable length = 6.54), and each phrase contained 2 pitch accents of the same type (either two H*s or two L+H*s), and all ended in the same boundary tone (L-L%). The selection of phrases with two similar pitch accents and an L-L% was made due to the fact that it the stimuli came from naturalistic speech, in which it is difficult to control a priori for the number of pitch accents a speaker may use. Indeed, the combination of two of the same pitch accents and an L-L% boundary tone was the only possible one that was testable given the parameters of the original corpus. As the tokens were extracted from casual speech recordings, it was not possible to control entirely for the semantic content of the phrases, and each phrase was uttered spontaneously in the conversational task context. However, care was taken to avoid tokens that contained explicit discussions of race as well as lexical items that might be associated with AAL (phrases with words such as “dope” and “homie”, which appeared in the corpus, were not included in the stimuli, for example). Additionally, Author (2019) explored the use of 40 segmental phonological and morphosyntactic features of AAL in the same corpus, and found that they occurred extremely infrequently (mean N = <8 occurrences/20 min task), thus somewhat mitigating possible effects of these other types of features. Of course, there are a number of additional phonetic features that may correlate with AAE which could influence such judgments, though they were beyond the scope of the current analysis. Though the use of casual speech creates some unique limitations, it has the advantage of being naturalistic and therefore may be more appropriate for testing how listeners may make ethnicity judgments in realistic situations.

With respect to focus, Author (2016) labelled the phrases in the corpus as having broad or narrow focus, based on syntactic, semantic, and phonetic criteria. The phrases tested in this experiment are taken directly from that data set, with their corresponding focus labels. The combinations of pitch accents and focus marking in the stimuli, along with example phrases, appear in the **Table 1** below. This table contains all of the combinations of variables of interest that the listeners heard from one speaker in order to provide the reader with greater clarity about the experimental design.

TABLE 1 | Stimuli set for one speaker with 8 clips under the 4 different possible intonational conditions. In narrow focus conditions, the word where the narrow focus appears is indicated in bold. Example phrases from one participant with narrow focus lexical items in bold.

Clip	PA Type	Focus Type
of thirty-one years	H*	B
livin’ in that house	H*	B
if she went to school or not	H*	N
then I think I would	H*	N
and even before that	L+H*	B
four days off	L+H*	B
and it was a union job	L+H*	N
thing that I can imagine	L+H*	N

The experiment was designed in this fashion in order to allow of number of different direct comparisons during the analysis phase. These comparisons were as follows:

1. The effect of H* vs. L+H* regardless of focus.
2. The effect of broad vs. narrow focus, regardless of pitch accent type.
3. The effect of low-pass filtering vs. original clips on ratings, independent of intonational contours and focus.
4. The effect of low-pass filtering vs. original clips on ratings as a result of broad vs. narrow focus and/or H* vs. L+H* pitch accent types.

a. The Experiment

45 listeners were recruited via a university participant pool and well as through friend-of-a-friend methods. The listeners all identified as Black or White and as male or female, and the sample was balanced to obtain comparable numbers of listeners from each gender/race pairing[5]. Listeners were primarily undergraduates at a large, private university and the experiment was conducted in a quiet room in a university’s phonetics laboratory. Upon arrival, listeners were instructed that the experiment would proceed in two parts. Listeners were outfitted with a pair of Bose headphones, and then followed the experiment in an online survey hosted by Qualtrics. They read and agreed to a consent form and then heard a sample sentence for which they were asked to decide whether the speaker was Black or White. After that, they began experiment Task 1. In Task 1, listeners heard 48 clips in a randomized order (8 from each speaker, and counterbalanced for focus and pitch accent type variables and following each clip), and were asked to respond to the binary choice question “What is the ethnicity of the speaker” as quickly as possible. In Task 1, the clips were low pass filtered at 400 Hz, following Knoll et al., 2009, in order to obscure most segmental information but retain F0

[5] The sample contained 10 White men, 10 White women, 10 Black men, and 15 Black women. Participants from other racial/ethnic groups were excluded due to a lack of clarity in earlier literature about the perceptions of MUSE and AAL among groups of non-Black and non-White individuals.

information (Thomas and Reaser 2004). Listeners were instructed to open the door and alert the researcher when they completed Task 1. The researcher then entered the room and confirmed that the participant had reached the end of Task 1.

In Task 2, the listeners repeated the same task, but this time they heard the original unmodified versions of the same phrases. Following Tasks 1 and 2, the listeners were asked the following series of questions about the experiment:

1. What did you think of the tasks?
2. How easy or difficult or easy was the task?
3. How many different voices do you think you heard in each part?
4. Do you have any other comments about the experiment?

Following these questions, the participants were also asked a series of open-ended demographic questions including their gender, age, level of education, race, places of residence during their lifetime, and linguistic ability in languages other than English. This data was examined qualitatively to check for broad patterns related to listener experience. In the end, participants were age 20–30 and either current university students or recent graduates, so age and education were not variable enough to test for listener differences. From the qualitative analysis, which was necessary due to the small data set, there were also no clear patterns with respect to region or L1 experience. The regression models do include gender and race as factors however, since these were the only factors that could be included in the model and still yield functional results.

ANALYSIS

The analytical methods employed were designed to examine whether pitch accent type, focus type, or their interaction affected listener judgments of ethnicity by comparing clips in which these factors varied reliably. They were also designed to control for aspects of listener demographics, such as gender and race and the interaction thereof. Multiple logistic regression models were conducted in R using the lme4 (Bates et al., 2015) package and plotted using ggplot2 (Wickham 2009), and the results of the regression models are presented in the tables in **Appendix A** (Task 1) and **Appendix B** (Task 2). Both models presented here controlled for the effects of speaker and subject as random effects in order to mitigate the effects of individual variation, and since each speaker uttered different tokens, the effects of utterance are also partially controlled, though it was impossible to include token as a random effect due to the relatively small size of the stimuli set. The results for the clips in Task 1 and Task 2 appear to differ substantially, as well as indicate that speakers may be using a different decision-making process for these clips. As a result, the analysis for these two tasks will be presented independently, with discussion and comparison of the two tasks to follow.

Task 1: Filtered Clips

In Task 1, listeners were presented with a Qualtrics survey that contained 48 clips filtered at 400 Hz, building on the methods discussed in Thomas and Reaser (2004) and Knoll et al. (2009). A filtering condition of 400 Hz was chosen to maintain features of F0 while obscuring the majority of formant and segmental information. These clips were counterbalanced for the variables of pitch accent type, focus type, and interaction, as presented in **Table 1** above. Listeners were presented with each clip and then instructed to respond to the forced choice question "What is the ethnicity of this speaker?" and given the options of "Black" or "White". They were instructed to respond to this question as quickly as possible.

The logistic regression model fitted for this task was (Response~SubjectGender*SubjectRace+PA*Focus+(1|Speaker)+(1|Subject), family=binomial). Results of this model examining listener responses with pitch accent type, focus type, listener gender, and listener race as fixed effects, and speaker and listener as random effects reveal a significant main effect for pitch accent type, indicating that the stimuli with the L+H* pitch accents were significantly more likely to be labeled as having been uttered by a Black speaker ($p < 0.001$), as can be observed in **Figure 3**. In contrast, the results reveal no significant main effect

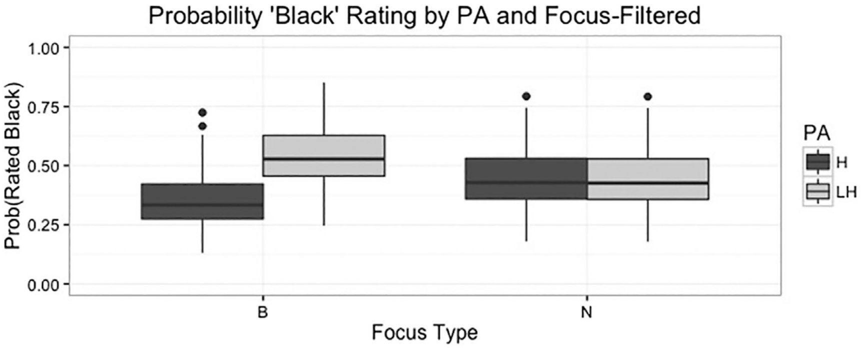


FIGURE 3 | Boxplot showing likelihood of stimuli being rated 'Black' based on pitch accent ratings, focus type, and the interaction for the filtered listening condition. The bars on the left show that the combination of L+H* and broad focus is more likely to be rated as "Black" than the combination of H* and broad focus. The bars on the right show no difference between ratings of the two different pitch accent types in the narrow focus conditions.

of focus on likelihood of being rated as having been uttered by a Black speaker.

However, there is also an interaction between focus and pitch accent type, such that clips with the combination of broad focus and L+H* contours are more likely to be rated as Black ($p < 0.001$) though there is no significant difference between pitch accent types in the narrow focus condition. **Figure 3** shows these results for the main effects as well as this interaction.

As we can observe from **Figure 3** above, in the narrow focus condition on the right, pitch accent does not appear to affect the probability of a clip being rated "Black". However, in the broad focus condition on the left, clips with the L+H* pitch accent are significantly more likely to be rated as having been uttered by a Black speaker.

The regression model controlled for speaker and subject as random effects, in order to ensure that these observed differences were not primarily driven by the ratings of a particular speaker or listener subject. Finally, perhaps surprisingly, no significant results were obtained in the analysis of listener race/gender or the interaction of these variables, indicating that listener judgments in this task do not appear to be subject to variation based on those aspects of listener identity, in contrast with the results obtained in earlier studies by Foreman (2000) and Thomas and Reaser (2004).

Task 2: Unmodified Clips

In Task 2, which immediately followed Task 1, listeners were presented with another Qualtrics survey that contained 48 randomized clips, but in this condition, the clips were in their original, unmodified versions. These clips were identical to the clips in Task 1 except that they were unfiltered, and so were also counterbalanced for the variables of pitch accent type, focus type, and interaction, as discussed above. Listeners were again presented with each clip and then instructed to respond to the forced choice question "What is the ethnicity of this speaker?" and given the options of "Black" or "White". They were instructed to respond to this question as quickly as possible.

The logistic regression model fitted for this task was identical to the one fitted for task 1, except for now it was run on responses unfiltered stimuli. The formula was Response~SubjectGender*SubjectRace+PA*Focus+(1|Speaker)+(1|Subject), family=binomial. Results of this logistic regression model examining listener responses with focus type, pitch accent type, subject gender, and subject race as fixed effects, and speaker and listener as random effects reveal that pitch accent type is a significant factor such that clips with the L+H* are less likely to be labeled as having been uttered by a Black speaker ($p < 0.001$). With respect to the question of focus, we also obtain a significant effect such that stimuli with the narrow focus were also less likely to be labeled as Black ($p = 0.00273$). **Figure 4** shows these results.

These results represent somewhat of a reversal of the trends observed for the responses in Task 1, though unlike in Task 1, the interaction of the two variables was not significant. Overall, broad focus tokens were less likely to be rated as "Black", as were those with the L+H* pitch accent. The regression model again controlled for the effects of speaker and subject, in order to ensure that these observed differences were not primarily driven by the ratings of a particular speaker or listener subject. Again, no significant differences were observed between groups of listeners organized by race, gender, or the interaction.

Comparing Tasks

In general, for the main effects, we can observe contrasting patterns for listener ratings between the filtered and original listening conditions, which is a result not previously documented in studies of ethnic identification. Additionally, we observe some interactions between pitch accent type and focus type with respect to the ratings. **Table 2A** below synthesizes the main effect results obtained in the previous two sections.

As we can observe from this table, the ratings (probability of rated "Black") pattern in opposite directions for the filtered and original listening conditions for the two different types of pitch accents. While focus does not appear to act as a main effect influencing ratings in the filtered condition, it is significant in the original condition. The table below synthesizes the interaction results obtained in the previous two sections.

As is evident from the results the **Table 2B** above, the effects pattern in opposite directions for the filtered vs. the original clips

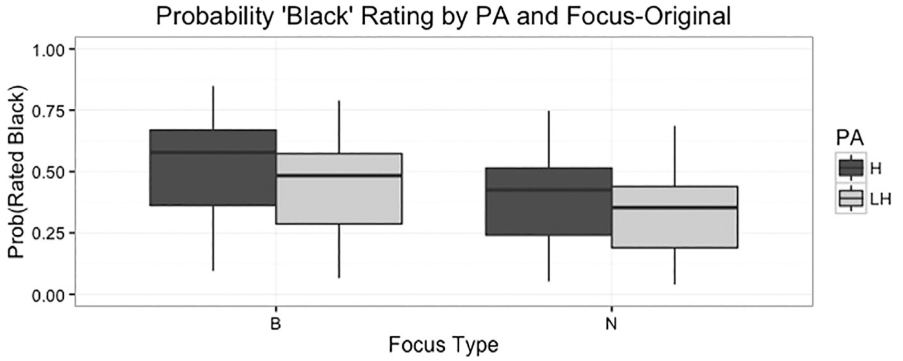


FIGURE 4 | Boxplot with error bars showing pitch accent ratings by focus type for the original stimuli listening condition. The bars on the left show that the H* accent is more likely to be rated as "Black" in the broad focus condition. The bars on the right show the same pattern for the narrow focus type condition, indicating no interaction of pitch accent and focus in this condition.

TABLE 2A | Results for main effects of pitch accent type as well as focus type in each listening condition.

Variable	Filtered Condition (judged as)	Original Condition (judged as)
H*	White	Black
L+H*	Black	White
Broad Focus	None	Black
Narrow Focus	None	White

TABLE 2B | Results for interaction effects of pitch accent type as well as focus type in each listening condition.

Interaction Variables	Filtered Condition (judged as)	Original Condition (judged as)
L+H*+Narrow	None	White
H*+Narrow	None	Black
L+H*+Broad	Black	White
H*+Broad	White	Black

with respect to the interaction of pitch accent and focus. This result stands in contrast to the results of other studies which have found that low-pass filtering causes listeners to be less accurate in their judgments or to simply behave at chance for filtered segments (Bryden 1968; Lass et al., 1980; Thomas and Reaser 2004).

Finally, with respect to potential task effects, it does not appear that listeners were subject to training effects of the tasks, though it is important to note that all participants heard the filter clips before the unmodified clips. This design was intended to prevent training effects, given that clips with more segmental information may be more easily recognizable that clips with less such information, though it does present a limitation of the current study since results can only be interpreted given this testing order. During the experiment debrief, listeners were asked to report the number of voices that they thought they heard across the two tasks. If listeners were indeed responding to a training effect, we may expect that they would report hearing a lower number of voices than what they actually heard. The mean number of voices reported for the listeners who was 8.31, though 7 out of the 43 listeners responded, "I don't know", indicating that over 16% of the sample was uncomfortable guessing how many voices were in the stimuli. This provides some evidence against a noticeable training effect, as does the fact that stimuli were randomized for each speaker in each task. Future studies, however, should consider additional methods for randomizing stimuli presentation in order to further test for such effects.

DISCUSSION

The results of this study which tested listeners' ratings of clips as "Black" or "White" under two listening conditions, original, and low pass filtered, while controlling for specific intonational phenomena of pitch accent type (L+H* vs. H*) and utilizing clips that had broad vs. narrow focus, yielded results that show that listeners appear to interpret these intonational phenomena in different, sometimes opposite, ways in filtered vs. original listening conditions. Previous studies on ethnic identification have generally found that listeners were less accurate under filtered conditions, but to date, this is the first study that has found that listeners may actually judge filtered and unfiltered clips in significantly different ways. By controlling for intonational contours as well as focus type, we have observed that listeners may differentially interpret the effects of intonational contours based on listening condition when attempting ethnic identification. This finding is particularly important for computational applications, especially ASR and speech synthesis systems, because if meaning of a particular contour is context-dependent for listeners, accurate systems must also take this into account. When attempting to accurately synthesize the speech of a Black speaker such that a listener would receive accurate sociolinguistic information, systems would necessarily have to account for how listeners make different judgments depending on acoustic quality. This is of particular importance since Black speakers have been historically underserved by computational systems; as Blodgett et al. (2020) note: "in the technology industry, speakers of AAE are often not considered consumers who matter" (9). This study thus provides important sociolinguistic context for computational researchers who aim to address this inequality.

This study also differs from previous studies in that its aim was not to test accuracy, but rather specifically examine differences between the effects of pitch accents, focus type, and filtering on listener judgments thus providing greater utility for computational applications. Given the results of McLarty (2018) and Author (2016) which have shown that AAL speakers may be more likely to employ the L+H* pitch accent than MUSE speakers, especially in broad focus contexts, one might expect that listeners would be consistently more likely to rate the L+H* pitch accent as having been uttered by a Black speaker. Additionally, (Author 2016), found that BWI speakers showed a pattern such that those who identified as more Black were more likely to use L+H*, and that this was especially the case in broad focus conditions. However, a consistent relationship between these patterns of production and perception was not obtained for the clips in the original stimuli in this study. Listeners appear more likely to judge that contour as "Black" only in low-pass filtered condition and not when they hear the original stimuli. These results also indicate that speakers' interpretation of intonational variables can differ depending primarily upon how much linguistic information they have available to them. That is, when speakers were exposed to filtered speech, hearing the L+H* pitch accent caused them to be more likely to rate the voice as Black. Interestingly, however, the interaction of narrow focus and L+H* gets rated as LESS Black, perhaps due to differences in salience and meaning of that pitch accent between MUSE and AAL (Thomas 2015, Author 2016). In particular, since earlier research has found that MUSE listeners may expect L+H* to signal contrastive meaning, it may be more marked in situations where it does not perform that function, such as in phrases with broad focus (Pierrehumbert and

Hirschberg 1990; Watson et al., 2008). Indeed, for both ASR and synthesis, capturing such ethnolinguistic different in prosodic contour meaning will be important for addressing not only user experience, but also bias in systems. Given that ASR is increasingly used for a variety of purposes, understanding whether or not a particular syllable has a pitch accent that signals contrastive focus may be important for the interpretation of the meaning of entire phrases.

Additionally, there is the possibility that the meaning of the L+H* pitch accent may differ between ethnolinguistic varieties, and also therefore potentially influence listener judgements. While the current study did not observe systematic differences with respect to judgments related to listener race, information about how individual listeners interpret the meanings of these intonational contours may further shed light on the mechanisms by which speakers make ethnicity judgments. As we have observed in earlier studies, especially Foreman (2000) and Thomas et al. (2010), Black and White listeners do sometimes pattern differently in ethnic identification tasks. Future studies should specifically address the potential for differential interpretations of the ethnolinguistic meanings of specific intonational contours for groups of listeners with different demographics. If the L+H* pitch accent sounds generally more marked/less standard for many listeners, synthesis systems must learn to employ pitch accents and boundary tones in a naturalistic way for voices that may be designed to represent different ethnolinguistic backgrounds.

Previous studies on ethnic identification have found that listeners may attune to a number of segmental and suprasegmental features in making ethnicity judgments, but that intonational variation does seem to play a significant role (cf Thomas et al., 2010). While the current study controlled for random effects of speaker and subject, it was unable to control for segmental phonological features due to the fact that it employed naturalistic speech. Intonation studies often face a difficult task of balancing the desire for control of segmental and syntactic information with the desire for naturalistic speech, and so it is possible that some features which were not entirely controlled for in the current study may also interact with the results obtained. Future work could compliment the results obtained here by using read speech, though that introduces a complication related to prosodic naturalness. However, comparing the results of studies that examined naturalistic vs. controlled speech might better shed light on these possible effects. The interaction between prosodic and segmental phonological variables will also be important for both ASR and speech synthesis systems, given that they also frequently rely on naturalistic speech. In particular, though systems may be improving in the naturalness of the production of segments, failure to replicate naturalistic prosody, or to combine naturalistic segments and contours will also limit improvements in speech synthesis.

These findings may also have broader consequences for linguistic profiling, which can have negative impacts on speakers' educational opportunities, economic prospects, as well as other types of interactions with government systems (Baugh 2003, 2015). Though the research has primarily focused on the ways in which stigmatized segmental and grammatical features may influence profiling, the fact that intonation is salient for listeners means that linguists need much more information on how listeners and speakers perceive and employ intonational variation at every level (Thomas 2015). Teaching speakers and listeners as well as communities to recognize the linguistic variables that may affect their perception of certain voices may be an important first step towards mitigating the often unconscious effects of linguistic profiling. With respect to ASR systems, a better understanding of the mechanisms by which this type of linguistic profiling occurs may also prevent future systems from miscategorizing or misevaluating the speech of user who employ non-standard varieties, thus creating a more equitable user experience. For speech synthesis, comprehensive models that rely on replicating naturalistic variation at all levels of linguistic structure will better serve individuals from a variety of background, thus improving their user experience and quality of life.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the New York University IRB. The patients/ participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4 : Linear Mixed-Effects Models Using S4 Classes. *R. Package*.

Baugh, J. (2003). “Linguistic Profiling,” in *Black Linguistics: Language, Society, and Politics in Africa and the Americas*. Editors S. Makoni, G. Smitherman, A. Ball, and A. Spears (New York: Routledge), 155–68.

Baugh, J. (2000). Racial Identification by Speech. *Am. Speech* 75 (4), 362–364. doi:10.1215/00031283-75-4-362

Baugh, J. (2015). “Speaking while Black.,” in *The Oxford Handbook of African American Language*. Editor S. Lanehart (UK: Oxford University Press).

Beckman, M. E., and Ayers, G. (1997). *Guidelines for ToBI Labelling*. Stillwater, OK: The OSU Research Foundation, 3.

Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. (2005). “The Original ToBI System and the Evolution of the ToBI Framework,” in *Prosodic Typology:*

The Phonology of Intonation and Phrasing. Editor S. A. Jun (UK: Oxford University Press), 9–54. doi:10.1093/acprof:oso/9780199249633.003.0002

Beckman, M. E., and Pierrehumbert, J. B. (1986). Intonational Structure in Japanese and English. *Phonol. Yearb.* 3 (01), 255–309. doi:10.1017/s095267570000066x

Blodgett, S. L., Barocas, S., Daumé, H., III, and Wallach, H. (2020). Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: 5454–5476.

Boersma, P., and Weenink, D. (2014). Praat: Doing phonetics by computer.

Bryden, J. D. (1968). The Effect of Signal Bandwidth Compression on Listener Perception. *J. Speech Hearing Assoc. Va.* 9, 6–13.

Cole, J., and Shattuck-Hufnagel, S. (2016). New Methods for Prosodic Transcription: Capturing Variability as a Source of Information: *J. Assoc. Lab. Phonology* 7 (1), 8. doi:10.5334/labphon.29

Crabtree, M., Mirenda, P., and Beukelman, D. (1990). Age and Gender Preferences for Synthetic and Natural Speech. *Augmentative Altern. Commun.* 6 (4), 256–261. doi:10.1080/07434619012331275544

Creer, S., Cunningham, S., Green, P., and Yamagishi, J. (2013). Building Personalised Synthetic Voices for Individuals with Severe Speech Impairment. *Comp. Speech Lang.* 27 (6), 1178–1193. doi:10.1016/j.csl.2012.10.001

Foreman, C. G. (2000). "Identification of African-American English from Prosodic Cues," in *Texas Linguistic Forum* (Austin: University of Texas), 43, 57–66.

Hlavac, M. (2015). *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. (R package version 5.2 http://CRAN.R-project.org/package=stargazer.

Holliday, N. R., and Jaggers, Z. S. (2015). Influence of Suprasegmental Features on Perceived Ethnicity of American Politicians. *Proc. 18th Int. Congress Phonetic Sci.*.

Jun, S.-A., and Forman, C. (1996). *Boundary Tones and Focus Realization in African American Intonation*. Paper presented at the 3rd Joint Meeting of the Acoustical Society of America and the Acoustical Society of JapanHonolulu

Knoll, M. A., Uther, M., and Costall., A. (2009). Effects of Low-Pass Filtering on the Judgment of Vocal Affect in Speech Directed to Infants, Adults and Foreigners. *Speech Commun.* 51 (3), 210–216. doi:10.1016/j.specom.2008.08.001

Koutstaal, C. W., and Jackson, F. L. (1971). Race Identifcation on the Basis of Biased Speech Samples. *Ohio J. Speech Hearing* 6, 48–51.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (1960). Evaluational Reactions to Spoken Languages. *J. Abnormal Soc. Psychol.* 60, 44–51. doi:10.1037/h0044430

Lanehart, S. (2015). *The Oxford Handbook of African American Language*. UK: Oxford University Press.

Lass, N. J., Almerino, C. A., Jordan, L. F., and Walsh, J. M. (1980). The Effect of Filtered Speech on Speaker Race and Sex Identifications. *J. Phonetics* 8, 101–112.

Lehr, M., Gorman, K., and Shafran, I., (2014). Discriminative Pronunciation Modeling for Dialectal Speech Recognition. In Fifteenth Annual Conference of the International Speech Communication Association.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An Overview of Noise-Robust Automatic Speech Recognition. *Ieee/acm Trans. Audio Speech Lang. Process.* 22 (4), 745–777. doi:10.1109/taslp.2014.2304637

McLarty, J. (2018). African American Language and European American English Intonation Variation over Time in the American South. *Am. Speech: A Q. Linguistic Usage* 93 (1), 32–78. doi:10.1215/00031283-6904032

Pierrehumbert, J. B. (1980). "The Phonology and Phonetics of English Intonation," in *Doctoral Dissertation* (Massachusetts: Massachusetts Institute of Technology).

Pierrehumbert, J., and Hirschberg, J. (1990). "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*. Editors P. R. Cohen, J. Morgan, and M. E. Pollack (Massachusetts: MIT Press), 271–311.

Purnell, T., Idsardi, W., and Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *J. Lang. Soc. Psychol.* 18 (1), 10–30. doi:10.1177/0261927x99018001002

Rahman, J. (2008). Middle-Class African Americans: Reactions and Attitudes toward African American English. *Am. Speech* 83 (2), 141–176. doi:10.1215/00031283-2008-009

Rockquemore, K., Brunsma, D. L., and Feagin, J. R. (2008). *Beyond Black: Biracial Identity in America*. Maryland: Rowman & Littlefield.

Shah, D. S., Schwartz, H. A., and Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Pennsylvania, United States. 5248–5264.

Shriberg, E., and Stolcke, A. (2004). "Prosody Modeling for Automatic Speech Recognition and Understanding," in *Mathematical Foundations of Speech and Language Processing* (New York, NY: Springer), 105–114. doi:10.1007/978-1-4419-9017-4_5

Spears, A. (1988). "Black American English," in *Anthropology for the Nineties: Introductory Readings*. Editor J. B. Cole (New York: Free Press), 96–113.

Tatman, R., and Kasten, C., (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In Interspeech. 934–938.

Thomas, E. (2015). "Prosodic Features of African American English," in *The Oxford Handbook of African American Language*. Editor S. Lanehart (UK: Oxford University Press), 420–438.

Thomas, E. R., Lass, N. J., and Carpenter, J. (2010). Identification of African American Speech, in *A Reader In Sociophonetics. Trends in Linguistics: Studies and Monographs 219*. Editors D. R. Preston and N. Niedzielski (New York: De Gruyter Mouton), 265–285.

Thomas, E. R., and Reaser., J. (2004). Delimiting Perceptual Cues Used for the Ethnic Labeling of African American and European American Voices. *J. Sociolinguistics* 8 (1), 54–87. doi:10.1111/j.1467-9841.2004.00251.x

Vicsi, K., and Szaszák, G. (2010). Using Prosody to Improve Automatic Speech Recognition. *Speech Commun.* 52 (5), 413–426. doi:10.1016/j.specom.2010.01.003

Watson, D. G., Tanenhaus, M. K., and Gunlogson, C. A. (2008). Interpreting Pitch Accents in Online Comprehension: H* vs. L+H*. *Cogn. Sci.* 32 (7), 1232–1244. doi:10.1080/03640210802138755

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. doi:10.1007/978-0-387-98141-3

Xu, Y., and Xu, C. X. (2005). Phonetic Realization of Focus in English Declarative Intonation. *J. Phonetics* 33 (2), 159–197. doi:10.1016/j.wocn.2004.11.001

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

A. Full Text of Stimuli Sentences.

Speaker	Clip	PA Type	Focus Type
1	of thirty-one years	H*	B
1	livin' in that house	H*	B
1	if she went to school or not	H*	N
1	then I think I would	H*	N
1	and even before that	L+H*	B
1	four days off	L+H*	B
1	and it was a union job	L+H*	N
1	thing that I can imagine	L+H*	N
2	to do something	H*	B
2	out of my mind	H*	B
2	at the store	H*	N
2	the second thing	H*	N
2	in my headphones	L+H*	B
2	after a while	L+H*	B
2	I was terrified	L+H*	N
2	now this is getting personal	L+H*	N
3	I thought were pretty	H*	B
3	being around	H*	B
3	terrified of heights	H*	N
3	you have to pass	H*	N
3	I would have	L+H*	B
3	really understand	L+H*	B
3	the college mentality	L+H*	N
3	in the long run	L+H*	N
4	close at like eleven	H*	B
4	that movie actually	H*	B
4	that you talk to girls	H*	N
4	that's why I hated going to work	H*	N
4	I don't really	L+H*	B
4	it was really cold	L+H*	B
4	have a job fair	L+H*	N
4	and even more fun	L+H*	N
5	I would say	H*	B
5	but not empowered in	H*	B
5	all these companies that	H*	N
5	being in the park	H*	N
5	and there you go	L+H*	B
5	come to africa	L+H*	B
5	in the bank	L+H*	N
5	after the show	L+H*	N
6	because of just	H*	B
6	I had a great time	H*	B
6	I did nothing	H*	N
6	getting attacked	H*	N
6	this woman got	L+H*	B
6	I don't know	L+H*	B
6	that's a good one	L+H*	N
6	three in the morning	L+H*	N

B. Regression Table for Task 1 (Filtered Clips) With Black/White as Response Variable, plotted using stargazer package (Hlavac 2015)

	Dependent variable:
	Response
SubjectGenderM	−0.216
	(0.191)
SubjectRaceW	0.091
	(0.192)
PALH	−0.779***
	(0.126)
FocusN	−0.379***
	(0.126)
SubjectGenderM:SubjectRaceW	−0.410
	(0.285)
PALH:FocusN	0.787***
	(0.178)
Constant	0.749***
	(0.216)
Observations	2,254
Log Likelihood	−1,474.723
Akaike Inf. Crit.	2,967.446
Bayesian Inf. Crit.	3,018.930

Note*: *p < 0.1;**p < 0.05; ***p < 0.01.*

C. Regression Table for Task 1 (Unmodified Clips) With Black/White as Response Variable, plotted using stargazer package (Hlavac 2015)

	Dependent variable
	Response
SubjectGenderM	−0.267
	(0.227)
SubjectRaceW	0.042
	(0.228)
PALH	0.405***
	(0.135)
FocusN	0.642***
	(0.136)
SubjectGenderM:SubjectRaceW	−0.212
	(0.339)
PALH:FocusN	−0.103
	(0.191)
Constant	0.085
	(0.359)
Observations	2,162
Log Likelihood	−1,311.175
Akaike Inf. Crit.	2,640.350
Bayesian Inf. Crit.	2,691.459

Note: **p < 0.1; **p < 0.05; ***p < 0.01.*

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 24 September 2021
doi: 10.3389/frai.2021.662097

Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems With DARLA

Rolando Coto-Solano, James N. Stanford* and Sravana K. Reddy**

Dartmouth College, Hanover, NH, United States

In recent decades, computational approaches to sociophonetic vowel analysis have been steadily increasing, and sociolinguists now frequently use semi-automated systems for phonetic alignment and vowel formant extraction, including FAVE (Forced Alignment and Vowel Extraction, Rosenfelder et al., 2011; Evanini et al., Proceedings of Interspeech, 2009), Penn Aligner (Yuan and Liberman, J. Acoust. Soc. America, 2008, 123, 3878), and DARLA (Dartmouth Linguistic Automation), (Reddy and Stanford, DARLA Dartmouth Linguistic Automation: Online Tools for Linguistic Research, 2015a). Yet these systems still have a major bottleneck: manual transcription. For most modern sociolinguistic vowel alignment and formant extraction, researchers must first create manual transcriptions. This human step is painstaking, time-consuming, and resource intensive. If this manual step could be replaced with completely automated methods, sociolinguists could potentially tap into vast datasets that have previously been unexplored, including legacy recordings that are underutilized due to lack of transcriptions. Moreover, if sociolinguists could quickly and accurately extract phonetic information from the millions of hours of new audio content posted on the Internet every day, a virtual ocean of speech from newly created podcasts, videos, live-streams, and other audio content would now inform research. How close are the current technological tools to achieving such groundbreaking changes for sociolinguistics? Prior work (Reddy et al., Proceedings of the North American Association for Computational Linguistics 2015 Conference, 2015b, 71–75) showed that an HMM-based Automated Speech Recognition system, trained with CMU Sphinx (Lamere et al., 2003), was accurate enough for DARLA to uncover evidence of the US Southern Vowel Shift without any human transcription. Even so, because that automatic speech recognition (ASR) system relied on a small training set, it produced numerous transcription errors. Six years have passed since that study, and since that time numerous end-to-end automatic speech recognition (ASR) algorithms have shown considerable improvement in transcription quality. One example of such a system is the RNN/CTC-based DeepSpeech from Mozilla (Hannun et al., 2014). (RNN stands for recurrent neural networks, the learning mechanism for DeepSpeech. CTC stands for connectionist temporal classification, the mechanism to merge phones into words). The present paper combines DeepSpeech with DARLA to push the technological envelope and determine how well contemporary ASR systems can perform in completely automated vowel analyses with sociolinguistic goals. Specifically, we used these techniques on audio recordings from 352

OPEN ACCESS

Edited by:
Joey Stanley,
Brigham Young University,
United States

Reviewed by:
Morgan Sonderegger,
McGill University, Canada
Eleanor Chodroff,
University of York, United Kingdom
Valerie Fridland,
University of Nevada, Reno,
United States

****Correspondence:***
Rolando Coto-Solano
Rolando.A.Coto.Solano@
Dartmouth.edu
James N. Stanford
James.N.Stanford@Dartmouth.edu
Sravana K. Reddy
sravana.reddy@gmail.com

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *31 January 2021*
Accepted: *26 August 2021*
Published: *24 September 2021*

Citation:
Coto-Solano R, Stanford JN and
Reddy SK (2021) Advances in
Completely Automated Vowel Analysis
for Sociophonetics: Using End-to-End
Speech Recognition Systems
With DARLA.
Front. Artif. Intell. 4:662097.
doi: 10.3389/frai.2021.662097

North American English speakers in the International Dialects of English Archive (IDEA[1]), extracting 88,500 tokens of vowels in stressed position from spontaneous, free speech passages. With this large dataset we conducted acoustic sociophonetic analyses of the Southern Vowel Shift and the Northern Cities Chain Shift in the North American IDEA speakers. We compared the results using three different sources of transcriptions: 1) IDEA's manual transcriptions as the baseline "ground truth", 2) the ASR built on CMU Sphinx used by Reddy et al. (Proceedings of the North American Association for Computational Linguistics 2015 Conference, 2015b, 71–75), and 3) the latest publicly available Mozilla DeepSpeech system. We input these three different transcriptions to DARLA, which automatically aligned and extracted the vowel formants from the 352 IDEA speakers. Our quantitative results show that newer ASR systems like DeepSpeech show considerable promise for sociolinguistic applications like DARLA. We found that DeepSpeech's automated transcriptions had significantly fewer character error rates than those from the prior Sphinx system (from 46 to 35%). When we performed the sociolinguistic analysis of the extracted vowel formants from DARLA, we found that the automated transcriptions from DeepSpeech matched the results from the ground truth for the Southern Vowel Shift (SVS): five vowels showed a shift in both transcriptions, and two vowels didn't show a shift in either transcription. The Northern Cities Shift (NCS) was more difficult to detect, but ground truth and DeepSpeech matched for four vowels: One of the vowels showed a clear shift, and three showed no shift in either transcription. Our study therefore shows how technology has made progress toward greater automation in vowel sociophonetics, while also showing what remains to be done. Our statistical modeling provides a quantified view of both the abilities and the limitations of a completely "hands-free" analysis of vowel shifts in a large dataset. Naturally, when comparing a completely automated system against a semi-automated system involving human manual work, there will always be a tradeoff between accuracy on the one hand versus speed and replicability on the other hand [Kendall and Joseph, Towards best practices in sociophonetics (with Marianna DiPaolo), 2014]. The amount of "noise" that can be tolerated for a given study will depend on the particular research goals and researchers' preferences. Nonetheless, our study shows that, for certain large-scale applications and research goals, a completely automated approach using publicly available ASR can produce meaningful sociolinguistic results across large datasets, and these results can be generated quickly, efficiently, and with full replicability.

Keywords: sociophonetics, vowels, dialects, American English, automated speech recognition, linguistics, Northern cities vowel shift, Southern vowel shift

INTRODUCTION

Phonetic alignment and extraction of vowel formants are central to modern sociophonetics (Thomas, 2011; Kendall and Fridland, 2021), and recent decades have seen a steady increase in automation for these important tasks. The FAVE system, Forced Alignment, and Vowel Extraction (Rosenfelder et al., 2011) provided one such semi-automated tool. With FAVE, users manually transcribe the text in Praat TextGrids (Boersma and Weenink 2019), upload to an automatic aligner (FAVE-Align), then use FAVE-Extract to extract the vowel formant frequencies. This produces an important improvement in processing time: Labov et al. (2013) report that, with 40 h of manual work his team could process the phonetic information of 300 vowels. On the other hand, using automatic alignment, up to 9,000 vowels could be processed in the same 40 h. But despite this progress, the current state-of-the-art methods still have to deal with an expensive and time-consuming bottleneck: the manual transcription of recordings. For accurate results, human transcribers must manually transcribe the audio. In this respect, most modern sociophonetic tools are "semi-automated," in that they require human transcription (or at least human verification of a transcription) to then proceed to the automated extraction of phonetic information. This step of manual transcription takes an

[1] Recordings and associated text files are copyright and used by special arrangement with the International Dialects of English Archive at https://www.dialectsarchive.com

enormous amount of time and resources of human labor, and frequently introduces human error due to typographical errors or other problems during annotation.

The DARLA system, which is short for "Dartmouth Linguistic Automation" (darla.dartmouth.edu) (Reddy and Stanford (2015a-c), provides a user-friendly version of this workflow which has become prevalent in recent years with researchers and students around the world; over 25,000 jobs have been run on DARLA since 2015. DARLA has a web-based utility for simple uploads of transcriptions (TextGrids or plaintext) and audio. Unlike other systems, DARLA has both a semi-automated and a fully automated system for vowel alignment and extraction. Both systems use the Montreal Forced Aligner for the phonetic alignment (McAuliffe et al., 2017). In the semi-automated version, users manually transcribe the audio into either plaintext files or audio-aligned TextGrids. In the fully automated system, users upload audio and DARLA uses its own in-house automatic speech recognition system (ASR) to create a transcription. After the ASR process is complete, DARLA goes on to align and extract the vowel formants, matching the audio and transcription with the Montreal Forced Aligner and then extracting the formants using FAVE-Extract (Rosenfelder et al., 2011). DARLA's current ASR is based on the CMU Sphinx toolkit, a HMM/GMM based ASR system. (HMM/GMM stands for "Hidden Markov Model, Gaussian Mixture Model," the mechanism for finding the phones in the audio stream). Reddy and Stanford (2015c) show that DARLA's fully automated transcription function can generate useful sociolinguistic results in a completely "hands-free" manner. The study used DARLA to automatically analyze US Southern and US Northern speakers, finding that the fully automated system could uncover statistically significant contrasts between the two regions in terms of the Southern Vowel Shift. Although these North-South contrasts were more clearly visible in the manually transcribed version, Reddy and Stanford (2015c) pointed out that despite limitations of the current ASR system, that fully automated system could still produce useful sociolinguistic results from some types of large-scale "big data" applications.

As examples a-d (reprinted from Reddy and Stanford, 2015c) below suggest, errors in the transcription may not affect the overall goal of producing vowel formants that are generally representative of a speaker's dialect features. In these examples, the ASR system has made large errors in transcription which crucially affect the meaning of some of the sentences. But from the sociophonetician's viewpoint, these errors may not affect the end result. In many cases, the extracted (stressed) vowel is the same for both systems, such as in the word those versus close and in spend versus depend. Naturally, phonetic environments may be affected [(z) in those versus (s) in close]. But for some large-scale applications, this may not be crucial. Reddy and Stanford (2015c) find that the US Southern Vowel Shift can be effectively diagnosed using such fully automated functions. Using 46 Southern and 47 Northern speakers in the Switchboard corpus (Godfrey and Holliman, 1993), they show a statistically significant difference between Southern speakers and Northern speakers, and they do this without needing any manual human transcribers.

a) *Manual*: give me your first impression. ASR: give me yours first impression
b) *Manual*: It's one of those. ASR: It's close
c) *Manual*: no It's It's wood turning. ASR: no it would turn it
d) *Manual*: and we really Don't spend on anything. ASR: and we don't depend on anything

Even though the fully automated pipeline has been shown to recover dialect differences, in practice, most users of DARLA depend on the semi-automated version with manual transcripts, since the word error rate of the CMU Sphinx ASR remains high.

In recent years, there have been numerous improvements in automatic speech recognition over the traditional HMM/GMM models. Two of them stand out: more availability of audio training data, and more powerful end-to-end deep learning algorithms. The amount of high-quality transcribed speech has exploded in recent years, and much of this data is available under open licenses. Two examples of such datasets are Mozilla's crowd-sourced Common Voice (Ardila et al., 2019), which contains more than 1,600 h of English, where volunteers read elicited sentences, and the OpenASR's LibriSpeech (Panayotov et al., 2015), which contains over 1,000 h of volunteers reading book passages. This has greatly increased the training data available to ASR algorithms, which themselves have improved during the last decade. The adoption of end-to-end algorithms has led to important reductions in transcription errors. These algorithms learn the word order and the phone acoustics together, rather than through separate language and acoustic models. They also build upon massive advances in deep learning, particularly in the capacity of their neural networks to understand the context of a word. The DeepSpeech algorithm (Hannun et al., 2014) uses these open corpora and combines them with an end-to-end architecture.

Our objective in this paper is to measure whether these end-to-end algorithms can provide transcriptions that are good enough to detect well-known sociophonetic patterns. There is research indicating that current ASR systems do not perform equally well with non-standard dialects of English (Tatman, 2017), so it is possible that such an experiment will fail to detect patterns such as the movements of vowels in the Southern dialect of US English. We conduct a test with speakers from all states and examine two regional dialects of US English: Southern and Inland North.

In the next sections we tackle the following questions: 1) Can automated transcriptions detect large-scale sociolinguistic patterns in a large dataset? 2) Can newer systems like DeepSpeech detect these patterns better than previous automated transcription methods? 3) How does an automated transcription fare against human-transcribed data in detecting sociolinguistic patterns? The first two questions will be studied in *Improvements in Sociophonetic Analysis* and *Variationist Analysis of NCS Movements*, and the third question will be studied in *Improvements in Sociophonetic Analysis* for the Southern Vowel Shift, and *Variationist Analysis of NCS Movements* for the Northern Cities Shift.

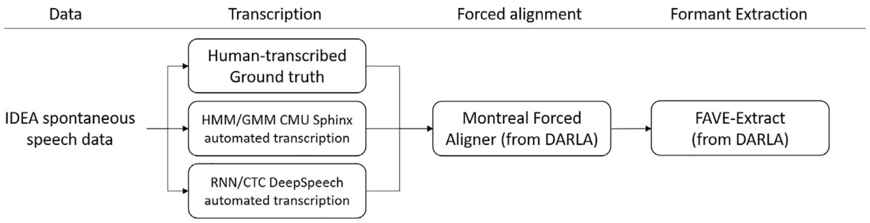

Data
Transcription
Forced alignment
Formant Extraction
IDEA spontaneous speech data
Human-transcribed Ground truth
HMM/GMM CMU Sphinx automated transcription
RNN/CTC DeepSpeech automated transcription
Montreal Forced Aligner (from DARLA)
FAVE-Extract (from DARLA)

FIGURE 1 | Workflow for data processing.

METHODS

We produced three types of transcriptions: (i) the manually transcribed ground truth, based on the IDEA transcription but hand-corrected to ensure it matches the recordings (ii) an automated transcription produced by the DeepSpeech program, (iii) and an automated transcription using the previously existing DARLA system built upon CMU Sphinx (Lamere et al., 2003). DeepSpeech was used with a pre-trained model developed by Mozilla[2], which is trained on the Fisher, LibriSpeech, Switchboard, and Common Voice English corpora in addition to 1700 h of transcribed NPR radio shows. The Sphinx system uses a model pretrained on a variety of American English speech corpora, mainly broadcast news and telephone conversations. **Figure 1** shows a summary of the workflow for data extraction.

These transcriptions were produced for recordings of 352 speakers of American English included in the International Dialects of English Archive[3] corpus (IDEA). All of the recordings are in a conversational, informal style, recorded in interviews asking the participants to talk about where they are from. There is only one recording per speaker, and, in total, the corpus contains approximately 12.5 h of audio (45,154 s). The recordings were an average of 128 ± 59 s long, with a minimum of 23 s and a maximum of 6 min 28 s. The corpus included 192 female and 160 male speakers, 54 and 46% respectively. The ages of the speakers at the time of recording ranged between 11 and 95 years old at the time of recording, with a median age of 37. The ethnic makeup of the sample is as follows: 79% was white (279), 10% was black (34), 5% was of Latin American descent (16), 3% was Native American (10), 0.28% was Asian American (1 person), 3% reported mixed ancestry 9), and 1% declared no ethnicity (3).

The data included speakers from every state in the United States. These speakers were grouped in three groups: Inland North, Southern, and General North. These three regional groupings made it possible for us to make regional comparisons of speakers in terms of the Southern Vowel Shift (SVS) and the Northern Cities Shift (NCS), as discussed below.

The Inland North group was defined according to the region identified as Inland North in the Atlas of North American English (ANAE) (Labov et al., 2006), as reprinted in **Figure 2**. In **Figure 2**, Our Inland North group was defined according to the region identified as Inland North in the Atlas of North American English (ANAE) (Labov et al., 2006, see ANAE page 148 map 11.15). The Inland North is the region around the US Great Lakes states and stretching east into New York state and also stretching downward along the "St. Louis Corridor" to St. Louis, following the ANAE analysis of this region as the Northern Cities Shift region. The Southern group was defined as speakers located in the traditional US South in the ANAE, not including Florida. Florida is exceptional since it has large amounts of immigration from northern US regions, and it has a different sociolinguistic history (controlled by Spain for a long period of time in the colonial era). Finally, our General North group was defined as all speakers not in the South, not in Florida, and not in the Inland North (and therefore, as roughly equivalent to Standard American English). This also includes Western varieties of American English. The reason for this analytical choice to define the General North broadly is that this broad region is known to contrast sharply both with the South and with the Inland North, as defined in the ANAE, in terms of two major vowels shifts considered here: The Southern Vowel Shift and the Northern Cities Shift. That is, in prior work (ANAE, Labov et al., 2006) the SVS vowel features were found in the South as defined here, and speakers in this region contrasted with speakers elsewhere in North America. Therefore, for SVS we compare speakers in the South group versus speakers in the General North group. As for NCS, the ANAE determined that the NCS vowel shift was found in the Inland North in contrast to the vowel system of the General North; the regional boundaries of the Inland North are defined in the ANAE in terms of this vowel shift that differs from the General North. Likewise, our NCS analysis compares Inland North speakers with General North speakers. In this way, we are able to test whether the NCS vowel contrast that the ANAE reported in terms of Inland North versus General North, which was based on manual vowel extractions, is also present in the IDEA data set using the automated methods of our present paper.

Once the recordings are transcribed, we calculated the Character Error Rate between (i) the ground truth transcription and the DeepSpeech automatic transcription and (ii) the ground truth and the CMU Sphinx transcription (see *Character Error Rate* below). We then extracted the formants of

[2] https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3

[3] Recordings and associated text files are copyright and used by special arrangement with the International Dialects of English Archive at https://www.dialectsarchive.com

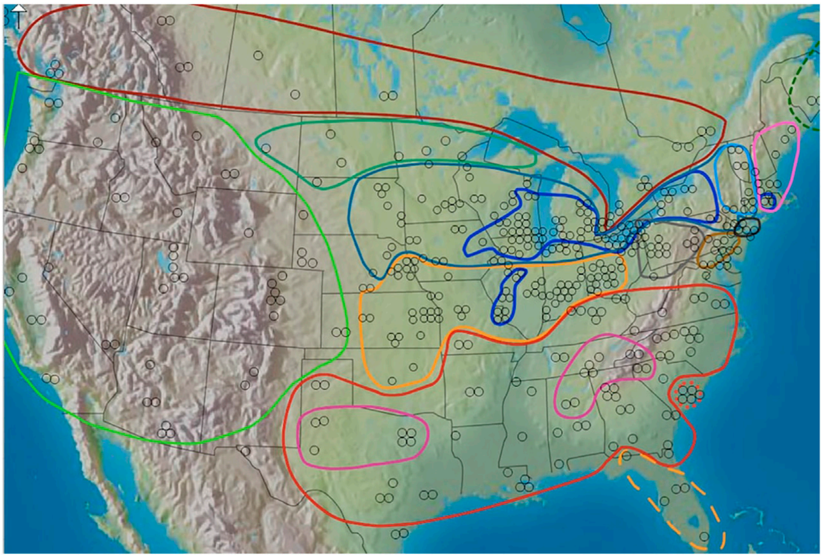


FIGURE 2 | North American dialect regions as outlined in the ANAE. Dark blue = Inland North. Red = South. Map to be reprinted from Labov et al. (2006).

the vowels in each transcription system (ground truth, Sphinx, and DeepSpeech) with the DARLA semi-automated system, which uses the Montreal Forced Aligner and FAVE-Extract to calculate the formant information. This workflow is shown in **Figure 1** above. We then constructed vocalic triangles, diagrams of the position of different vowels along F1 and F2. We then compared the positions of different vowels according to F1 and F2 and measured the degree of overlap between the ground truth vowels and both the Sphinx and the DeepSpeech transcribed vowels. Finally, we used this information to observe two well-known phenomena in American English: Southern Vowel Shift and the Northern Cities Vowel Shift.

Following standard methods in American English sociophonetics, we removed tokens of vowels in unstressed and reduced syllables since such tokens do not accurately represent the vowels being studied here (Thomas, 2011). Likewise, we removed tokens of vowels in function words (e.g., common grammatical words like "the," "and," etc.), and also any tokens with large, unreliable formant bandwidths (greater than 300 Hz). This filtering of high-bandwidth tokens is a standard way of ensuring that the tokens used in the study are based on reliable Linear Predictive Coding, since their LPC formant estimations are likely to be less reliable at a high-bandwidth (Hofmann, 2014:110, 162, 196; Ladefoged, 2003: 117; Thomas, 2011:47). To reduce the effects on varying phonetic environments, we also removed tokens where vowels are in pre-liquid position, following standard practice for such shifts (Fridland and Bartlett, 2006; Nesbitt, 2018). Finally, since physiology and other factors can affect vocal tract length and vowel formants, we normalized the vowel formant measurements using the Lobanov method (Lobanov, 1971; Kendall and Thomas, 2010). The Lobanov normalization method has been one of the more commonly used approaches in sociophonetics spanning a large amount of time up to the present (e.g., Thomas, 2011; Fridland et al., 2014; Grama and Kennedy, 2019; Fridland and Kendall, 2019; D'Onofrio and Van Hofwegen, 2020; Nesbitt, 2021). We recognize that Barreda's perceptual analyses (Barreda, 2020, Barreda, 2021) suggest a log-based method rather than Lobanov, and future work may take that approach. However, the prior DARLA testing (Reddy and Stanford, 2015a; Reddy and Stanford, 2015b) that we are comparing in the present study used the Lobanov method, and we prefer a direct comparison between the results here and previous ones. We also note that the Lobanov normalization is included in the FAVE output spreadsheets, and so computational sociolinguistics readers will be familiar with this output. We also note that there are a large number of different vowel normalization practices and debates in sociolinguistics (see Thomas and Kendall, 2007 online NORM site for detailed discussion of five such methods). We decided to use one of the more commonly accepted methods at the present time, the Lobanov method, recognizing that every method has its own strengths and weaknesses.

RESULTS

This section compares the two transcription methods we used (Sphinx and DeepSpeech) in the following ways: (i) How well their transcriptions overlap with manual transcriptions (*Character Error Rate*), and (ii) how effective they are in detecting sociolinguistic phenomena such as the Southern Vowel Shift (*Sociophonetic Results From the Southern Vowel Shift*) and Inland North Cities Shift (*Northern Cities Shift Results*).

Character Error Rate

In order to investigate the differences in error rate between the transcription methods, we used a linear mixed effects model with character error rate as the dependent variable. Character error rate (henceforth CER) is the edit distance between two strings. For example, if the ground truth had the transcription "BAT" and the ASR produced the transcription "CAT," then the CER would be 0.33[4]. The CER was log-transformed to meet the assumptions of linear-mixed effects models. As for the independent variables, we used the type of transcription (DeepSpeech versus Sphinx, henceforth DS and SPH), the gender of the speakers, the geographic area (Inland North, Southern, and General North) the estimated year of birth (from 1915 to 2007), and the interaction between transcription type and gender. This interaction was included because research has shown that ASR systems perform systematically worse on female voices (Tatman and Kasten, 2017). All categorical variables were encoded using treatment coding; the reference level for each of them was the first one alphabetically (type of transcription: DeepSpeech, gender: female, area: Inland North). The numerical variable estimated year of birth was centered by calculating the z-score of the variable. Finally, the model included a random intercept for speakers.

The DeepSpeech ASR does produce a statistically significant improvement in transcription, and this improvement is greater for females than for males [transcription by gender interaction: $\beta_{Male:DS} = -0.13 \pm 0.03$, t (347) = −4.0, $p < 0.001$]. As for the males, there is a reduction of 0.09 units in the character error rate (from $CER_{SPH/M} = 0.46 \pm 0.11$ to $CER_{DS/M} = 0.37 \pm 0.17$). On the other hand, the improvement is greater when transcribing speech from females. In this case, the reduction in character error rate is 0.13 units (from $CER_{SPH/F} = 0.46 \pm 0.13$ to $CER_{DS/F} = 0.33 \pm 0.16$). This result is particularly important given the known issues with transcription of female speech, and our results suggest that deep-learning algorithms may be closing the gender gap in ASR performance. The main effects are also significant, confirming the direction of the interaction: When all other factors are held constant, there is a main effect for gender [$\beta_{Male} = 0.13 \pm 0.04$, t (492) = 3.3, $p < 0.001$]: Overall, the transcription for males has a higher error rate (CER = 0.41) than the transcription for females (CER = 0.39). Likewise, there is a main effect for transcription type [$\beta_{Sphinx} = 0.40 \pm 0.02$, t (347) = 18.0, $p < 0.00001$]: On average, when all other factors are held constant, the transcription for Sphinx (CER = 0.46) had more errors than the DeepSpeech transcription (CER = 0.35). These main results should be interpreted in light of the interaction: These main effect results agree with the more general result that Sphinx has more errors than DeepSpeech for female speakers. As for the other variables in the model, there are no significant differences in CER by region of the recording: Inland North: 0.40, South: 0.41, General North: 0.39 ($p_{InlandNorth/GeneralNorth} = 0.67$, $p_{South/GeneralNorth} = 0.15$). There are no significant differences in CER by estimated year of birth either ($p = 0.30$). Finally, the random intercept for speakers explains a sizable portion of the remaining variance in the regression ($var_{speaker} = 0.08$, $var_{residual} = 0.05$). In summary, the DeepSpeech ASR does provide improvements in transcription, particularly for speech from female speakers.

In general, the reduction in CER is an indication that the DeepSpeech transcriptions are closer to the original. The examples below show the improvements in the transcriptions of a speaker from the Inland North region, specifically from Minnesota. While words like "mom" are mistranscribed by both systems (as man and men respectively), the DeepSpeech transcription produced the correct vowel in "me"/"he," and correctly transcribed the words "that they do" and "so my."

GT: So, my mom and me came down here for the orientation that they do.

DS: so my man and he came down here for the orientation that they do (CER = 0.15).

Sphinx: follow men and i mean came down here for the orientation of the u (CER = 0.34).

The examples below show transcriptions for speakers from the South, from Alabama and Louisiana respectively. In the Alabama example, the DeepSpeech transcription is completely correct, but the Sphinx transcription has a few problems, including missing the pronoun "I" and mistranscribing "born in Jackson County" as going to act in canton. The Louisiana example shows an example of audio that was grossly mistranscribed by both systems. Even though they both make mistakes, the DeepSpeech system is closer to the original. For example, the stressed vowels in the words "growing" and going are the same (OW), whereas the Sphinx transcription has ground for those segments, which has the vowel AW[5].

Alabama:

GT: I was born in northeastern Alabama. I was born in Jackson County.

DS: I was born in north eastern alabama I was born in jackson county (CER = 0.08).

Sphinx: I was born in northeastern alabama was going to act in canton (CER = 0.28).

Louisiana:

GT: Growing up with my sister, I always felt like I got the short end of the stick.

DS: the going on with my sister always felt like so i got the short instink (CER = 0.33).

Sphinx: the ground and women does your always the white jacket distorted (CER = 0.73).

Given that there is a significant improvement in transcriptions, our next question is: Do these new transcriptions extend our capabilities to detect sociophonetic patterns in automatically transcribed data? We will test these

[4] CER is defined as (substitutions + deletions + insertions)/length of source. In the case of BAT/CAT, only one letter is substituted in a string of length three, and therefore CER = 1/3 = 0.33. If both the source and the target transcriptions are identical (e.g., BAT/BAT), then the CER is zero. If the transcription is wrong but has the same length, then CER = 1 (e.g., BAT/DOG). If the transcription is longer than the original, then the CER can be greater than one (e.g., cat/foxes, CER = 1.66)

[5] The vowels in this paper are transcribed using the ARPABET system for American English, as found in the CMU dictionary and in FAVE output spreadsheets: IY = FLEECE, UW = GOOSE, IH = KIT, EY = FACE, EH = DRESS, AE = TRAP, AW = MOUTH, AY = PRICE, AA = LOT, AO = THOUGHT, AH = STRUT, OW = GOAT, ER = NURSE, UH = FOOT, OY = CHOICE, and AHR = START

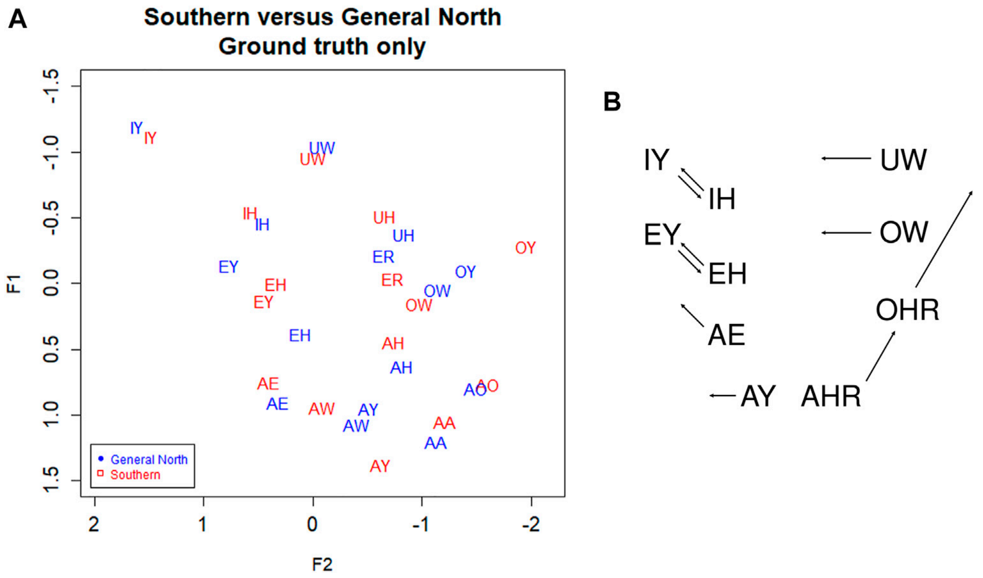


FIGURE 3 | **(A)** Southern versus General North data from manual transcription: Red = South, Blue = North. Ground truth (manual transcription). Speaker vowel means. 330 speakers, 31,900 tokens. Plotted in Lobanov-normalized units. **(B)** Abstract schematic of the Southern Vowel Shift (adapted from Labov, 1996).

by trying to observe the well-understood phenomena of two North American English vowel shifts: The Southern Vowel Shift and the Northern Cities Shift (Labov et al., 2006).

Sociophonetic Results From the Southern Vowel Shift

In the first subsection we will present the improvements in the sociophonetic analysis of speakers of Southern English. Following this, we will compare the Southern speakers with those of the General North, and analyze how the DeepSpeech transcription performs in comparing these two.

Improvements in Sociophonetic Analysis

In **Figure 3A**, we plot the Southern speakers versus the General North speakers using the ground truth (GT) transcription. We can see that the Southern Vowel Shift (SVS) is evident in this manually transcribed version of the data. First, note the EY/EH tense/lax shift in the General South (red) speakers such that EH becomes higher than EY (compare to the SVS schematic in **Figure 3B**). We also see some graphical evidence for the IY/IH tense/lax shift, although we expect this to be a weaker shift. Next, we note a highly advanced AW vowel in the General South speakers, as well as evidence of AE raising and OW-fronting. We note UW-fronting as well, but this is shared by both the General North and Southern speakers, suggesting an overall pattern of UW-fronting. We do not examine AXR and other complex shifts involving liquids since such movements go beyond the scope of the present study; likewise, the Southern monophthongization of AY and raising of OY and so on are topics for another study since they would require analysis of the off-glide.

Now consider **Figure 4B**, which shows Southern speakers in all three of the transcription types. **Figure 4A**, which includes the three transcription types for General North, is included for comparison. All three of the transcription types show the SVS features noted above, but ground truth (GT) and DeepSpeech (DS) show the clearest differences between the two dialects. In particular, note the configuration of EY/EH for the three transcription types. Both ground Truth and DeepSpeech show the full rotation as EH and EY "switch places" in the vowel space, as we would expect from the schematic in **Figure 3B**: the EY vowel (the vowel in FACE or MADE) retracts and lowers, while EH (the vowel in DRESS or RED) fronts and raises. By contrast, DARLA's current in-house Sphinx version only shows a general movement of EY/EH toward the SVS configuration but not the rotation. We expect that the Southerners' EY/EH shift will be more advanced than their IY/IH shift because this is commonly the case for the Southern Vowel Shift (Kendall and Fridland, 2012), and this is what we find in the figure. Moreover, we find that the DeepSpeech version more accurately reflects the status of the tense/lax shift than the Sphinx version.

We now examine each of the vowels in the SVS in terms of F1 and F2, comparing across all three transcription types: DeepSpeech, DARLA'S CMU Sphinx, and the ground truth. Following Johnson (2015) and Stanley (2018), we compare the token distributions using Bhattacharyya's Affinity (BA). This is calculated by describing each token by its two-dimensional coordinates (Lobanov-normalized F1 and F2), and then measuring the amount of overlap between the regions covered by both vowels. An affinity of 1.0 indicates a perfect overlap between the two distributions of vowel tokens, and an affinity of 0.0 indicates perfectly non-overlapping distributions. The formula and the concrete implementation used can be found in the kerneloverlap function in the R package adehabitatHR (R Core Team, 2021; Calenge, 2006). We use Bhattacharyya's Affinity rather than Pillai approaches since Johnson (2015) argues that BA

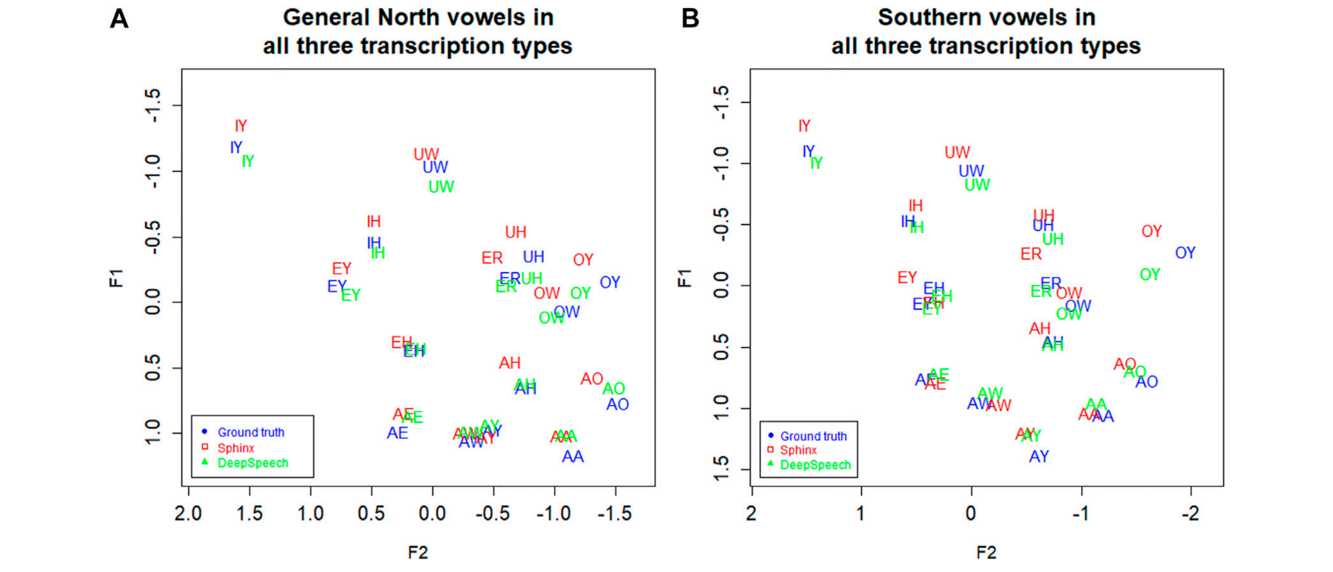


FIGURE 4 | (A) General North and **(B)** Southern speaker's vowels (mean position) in all three transcription types. Blue: Ground truth (manual) transcription, Red: CMU Sphinx transcription, Green: DeepSpeech transcription. 105 speakers and a total of 28,225 vowel tokens. Plotted in Lobanov-normalized units.

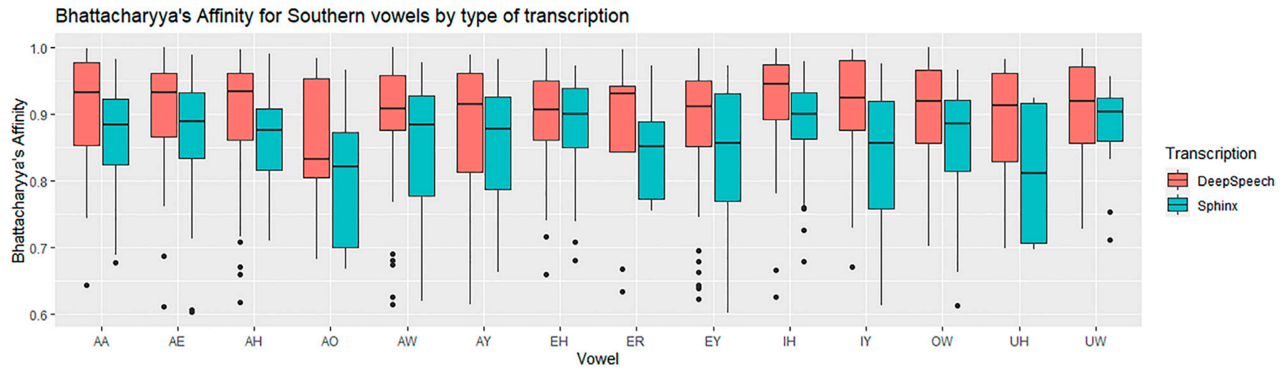


FIGURE 5 | Southern data: vowel medians for the Bhattacharyya's affinity by type of transcription. Red: DeepSpeech versus ground truth, Green = CMU Sphinx versus ground truth. 105 speakers and a total of 28,225 vowel tokens.

improves upon Pillai by more accurately quantifying overlap for the purposes of vowel distributions (for example, by better handling unequal distributions or distributions with an unequal number of tokens). In addition to this, Bhattacharyya's Affinity has been used to study vowel contrasts in New Zealand English (Warren, 2018) and in back vowels in Kansas (Strelluf 2016).

For each of the Southern Vowel Shift vowels, we compute the Bhattacharyya's Affinity for each speaker's distribution in terms of Sphinx versus ground truth, and then in terms of DeepSpeech versus ground truth. We then use a repeated-measures ANOVA to determine the relationship between Bhattacharyya's Affinity, type of transcription and the vowels in the transcripts. The affinity was used as the dependent variable, transformed with a reflected square root transformation to comply with normality assumptions. The vowels and the types of transcriptions were used as within-subjects independent variables.

There was a significant difference between the Sphinx transcription and the DeepSpeech transcription [F (1) = 25.8, $p < 0.00005$, $\eta^2 = 0.053$]. As can be seen in **Figure 5**, the vowels transcribed with DeepSpeech have a higher BA with the ground truth vowels. The median affinity for CMU Sphinx is 0.88, while the median affinity for DeepSpeech is 0.92. There was also a significant difference between vowels [F (13) = 1.2, $p < 0.00005$, $\eta^2 = 0.034$]: Some vowels have higher overall BAs (e.g., EH: 0.901, EY: 0.889, IH: 0.916, IY: 0.894), while others have significantly lower affinities (e.g., AO: 0.817). **Table 1** shows the vowels involved in the SVS. The interaction between vowels and type of transcription was not statistically significant (p = 0.26),

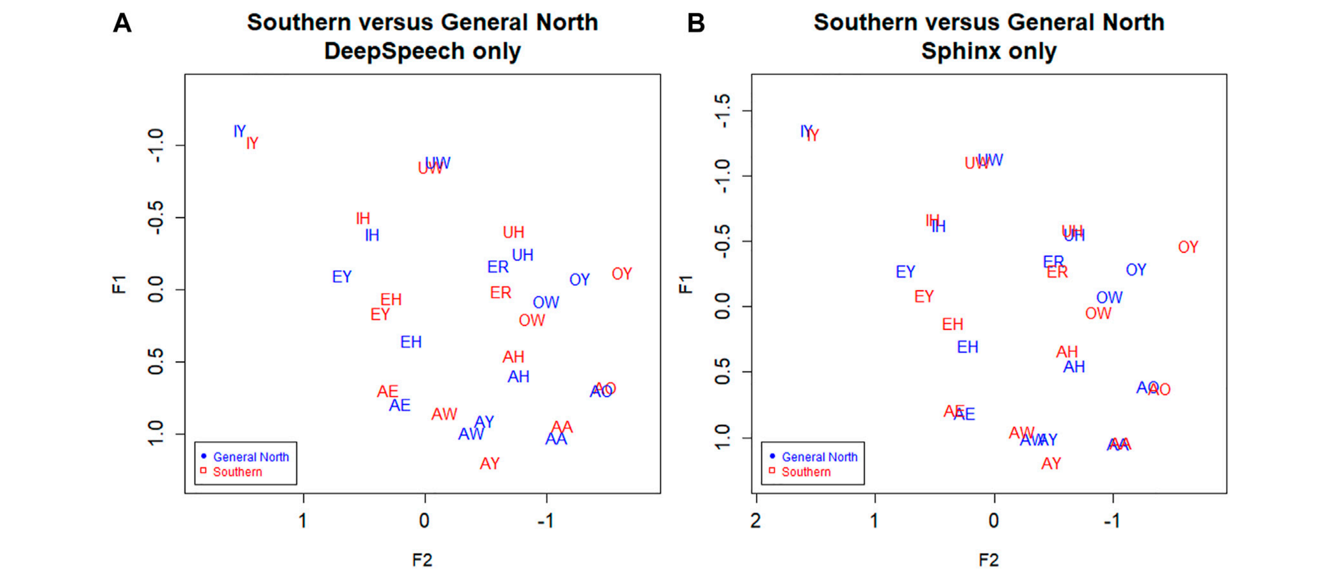


FIGURE 6 | Southern versus General North. Red: General North, Red: Southern. **(A)** DeepSpeech ASR, speaker vowel means. 330 speakers, 24,295 tokens. **(B)** CMU Sphinx ASR, speaker vowel means. 330 speakers, 27,637 tokens. Plotted in Lobanov-normalized units.

TABLE 1 | Bhattacharyya's Affinity for DeepSpeech versus ground truth and Sphinx versus ground truth in Southern speakers. BA score 1.0 = perfectly overlapping distributions, 0.0 = completely non-overlapping.

SVS vowel	Median BA for DS vs. GT	Median BA for SPH vs. GT	Δ(BA)
AW	0.907	0.883	0.024
EH	0.905	0.899	0.006
EY	0.910	0.855	0.055
IH	0.943	0.899	0.043
IY	0.923	0.850	0.074
OW	0.916	0.880	0.036
UW	0.918	0.903	0.015

meaning that no vowels were observed to have a marked improvement over others. In general, for all of the vowels involved in the SVS there is a gain in BA when transcribed automatically using DeepSpeech.

Variationist Analysis of SVS Movements

In the previous section we presented evidence that the DeepSpeech-based system is more effective than Sphinx at measuring the vowels of Southern speakers. Based on this, we can assume that the DeepSpeech system will also help in observing the vowel differences between Southern speakers and speakers of the General North variants. **Figure 6A** below shows the vowels from these two dialects, as extracted from the DeepSpeech data. Compare this to **Figure 6B**, the vowels as extracted by the Sphinx ASR system. The DeepSpeech system shows a clearer impressionistic separation between the two dialects. For example, the vowel IY shows a much clearer separation in the DeepSpeech data (**Figure 6A**), compared to the partial overlap in the Sphinx data (**Figure 6B**).

The next step is to conduct a variationist analysis of all of the major movements of the Southern Vowel Shift. Using linear mixed effects modeling with the lme4 package (Bates et al., 2015) from (R Core Team, 2021), we built models with the independent variables of Region (General North versus South), Year of Birth,[6] Gender, and Following Environment (nasal, voiceless obstruent, voiced obstruent). Year of birth is a numerical variable, so it was centered using z-scoring; the categorical variables were encoded using treatment coding. We also included the variable Transcription type (DeepSpeech versus ground truth), so that we can examine how well DeepSpeech holds up in comparison to the ground truth. The interaction between transcription type and region was also included, to determine whether the DeepSpeech automated transcription shows the North/South differences in a way that is, similar to the ground truth data (for example, by seeing of the degree of separation between Northern and Southern IY in the ground truth is also present in the DeepSpeech data). Finally, the election of the random effects for each model proceeded *via* backward selection from a maximal model, which included all variables (as well as the Region: Transcription interaction) for both speaker and word effects (Barr et al., 2013; Bates et al., 2015). We used the step instruction in (R Core Team, 2021), which gave us an optimal random effect structure for each of the vowels. The resulting models are shown in **Supplementary Appendix S1**.

The dependent variable for these models will vary according to the relevant variable for the motion of each vowel. For example,

[6] Most speakers in the IDEA dataset have a specific Year of Birth listed. But for a handful of speakers, the Year of Birth is just given as a decade, such as "1950s." For such speakers, we simply estimated the Year of Birth at the middle of the decade, i.e., 1955

TABLE 2 | LMER models for comparison between Southern Vowel Shift (SVS) and General North vowels. The dependent variable is F2 for the vowels MOUTH (AW), GOAT (OW), GOOSE (UW), and F2-2xF1 for the vowels TRAP (AE), FACE (EY), DRESS (EH), FLEECE (IY), and KIT (IH). R^2 shows marginal and conditional coefficient. Deltas show the difference between the mean position of the Northern vowel and the mean position of the Southern vowel for each transcription type: AE shows divergence in GT/DS results; for the other vowels, either both models detect a North/South difference or they do not.

A. Results for region and transcription type; R^2 for entire model

Vowels	Region by Transcription	Region	Transcription	Post-hoc ΔRegion_{GT}	Post-hoc ΔRegion_{DS}	R^2
AE (*n* = 5198)	$p = 0.08$	$p = 0.09$	$\beta_{GT} = -0.001 \pm 0.0004$ t (265) = −2.6, $p < 0.01$	$\Delta = 0.40$ z = −3.0, $p < 0.05$	$\Delta = 0.28$ $p = 0.33$	0.21 0.69
AW (*n* = 3116)	$\beta_{GT:South} = 0.02 \pm 0.01$ t (2567) = 2.2, $p < 0.05$	$\beta_{South} = 0.03 \pm 0.01$ t (179) = 2.9, $p < 0.005$	$\beta_{GT} = 0.03 \pm 0.008$ t (52) = 3.4, $p < 0.005$	$\Delta = 0.31$ z = −4.8, $p < 0.0001$	$\Delta = 0.22$ z = −2.9, $p < 0.05$	0.10 0.56
EY (*n* = 5730)	$p = 0.41$	$\beta_{South} = -0.07 \pm 0.01$ t (179) = −7.8, $p < 0.00001$	$\beta_{GT} = 0.02 \pm 0.004$ t (201) = 4.7, $p < 0.00001$	$\Delta = 0.86$ z = 7.9, $p < 0.0001$	$\Delta = 0.84$ z = 7.8, $p < 0.0001$	0.08 0.59
EH (*n* = 5930)	$\beta_{GT:South} = 0.02 \pm 0.01$ t (144) = 2.0, $p < 0.05$	$\beta_{South} = 0.06 \pm 0.01$ t (121) = −1.3, $p < 0.00001$	$p = 0.20$	$\Delta = 0.98$ z = −9.1, $p < 0.0001$	$\Delta = 0.75$ z = −6.2, $p < 0.0001$	0.10 0.53
IY (*n* = 4442)	$p = 0.86$	$p = 0.13$	$\beta_{GT} = 0.03 \pm 0.005$ t (134) = 7.2, $p < 0.00001$	$\Delta = 0.29$ $p = 0.15$	$\Delta = 0.26$ $p = 0.43$	0.08 0.55
IH (*n* = 5961)	$p = 0.58$	$\beta_{South} = 0.03 \pm 0.008$ t (282) = 4.0, $p < 0.00001$	$\beta_{GT} = 0.02 \pm 0.003$ t (333) = 5.2, $p < 0.00001$	$\Delta = 0.28$ z = −4.2, $p < 0.0005$	$\Delta = 0.30$ z = −4.0, $p < 0.0005$	0.02 0.62
OW (*n* = 4611)	$p = 0.25$	$\beta_{South} = 0.03 \pm 0.009$ t (250) = 3.2, $p < 0.005$	$\beta_{GT} = -0.03 \pm 0.004$ t (69) = −7.2, $p < 0.00001$	$\Delta = 0.16$ z = −4.5, $p < 0.0001$	$\Delta = 0.12$ z = −3.2, $p < 0.01$	0.03 0.50
UW (*n* = 2786)	$p = 0.70$	$\beta_{South} = 0.02 \pm 0.01$ t (255) = 2.1, $p < 0.05$	$p = 0.25$	$\Delta = 0.08$ $p = 0.07$	$\Delta = 0.06$ $p = 0.17$	0.01 0.56

B. Results for other social and linguistic variables in the model

Vowels	Year of birth	Gender	Following environment (Nasal versus voiced obstruent)	Following environment (Nasal versus voiceless obstruent)
AE (*n* = 5198)	$\beta = -0.001 \pm 0.0003$, t (195) = −2.3, $p < 0.05$	$p = 0.54$	$\beta_{Nas/+VoicedObs} = -0.12 \pm 0.01$, t (280) = −9.7, $p < 0.00001$	$\beta_{Nas/-VoicedObs} = -0.15 \pm 0.001$, t (525) = −15.7, $p < 0.00001$
AW (*n* = 3116)	$\beta = -0.02 \pm 0.004$, t (246) = −4.6, $p < 0.00001$	$p = 0.87$	$\beta_{Nas/+VoicedObs} = -0.06 \pm 0.01$, t (80) = −4.7, $p < 0.0001$	$\beta_{Nas/-VoicedObs} = -0.06 \pm 0.01$, t (108) = −4.5, $p < 0.0001$
EY (*n* = 5730)	$\beta = 0.02 \pm 0.004$, t (171) = 5.2, $p < 0.00001$	$p = 0.65$	$p = 0.34$	$\beta_{Nas/-VoicedObs} = 0.03 \pm 0.01$, t (196) = 2.4, $p < 0.05$
EH (*n* = 5930)	$\beta = -0.01 \pm 0.003$, t (283) = −3.2, $p < 0.005$	$\beta_{male} = 0.02 \pm 0.007$, t (283) = −3.2, $p < 0.0005$	$\beta_{Nas/+VoicedObs} = -0.03 \pm 0.01$, t (434) = −3.2, $p < 0.005$	$\beta_{Nas/-VoicedObs} = -0.06 \pm 0.01$, t (495) = −7.2, $p < 0.00001$
IY (*n* = 4442)	$\beta = 0.01 \pm 0.003$, t (257) = 4.3, $p < 0.00001$	$p = 0.15$	$\beta_{Nas/+VoicedObs} = 0.06 \pm 0.01$, t (115) = 5.4, $p < 0.00001$	$\beta_{Nas/-VoicedObs} = 0.08 \pm 0.01$, t (132) = 6.9, $p < 0.00001$
IH (*n* = 5961)	$\beta = 0.005 \pm 0.002$, t (286) = 2.2, $p < 0.05$	$p = 0.32$	$p = 0.17$	$\beta_{Nas/-VoicedObs} = 0.02 \pm 0.01$, t (424) = 2.2, $p < 0.05$
OW (*n* = 4611)	$\beta = 0.009 \pm 0.003$, t (297) = 2.7, $p < 0.01$	$p = 0.99$	$p = 0.30$	$p = 0.13$
UW (*n* = 2786)	$\beta = 0.01 \pm 0.003$, t (265) = 3.1, $p < 0.005$	$p = 0.79$	$p = 0.10$	$p = 0.76$

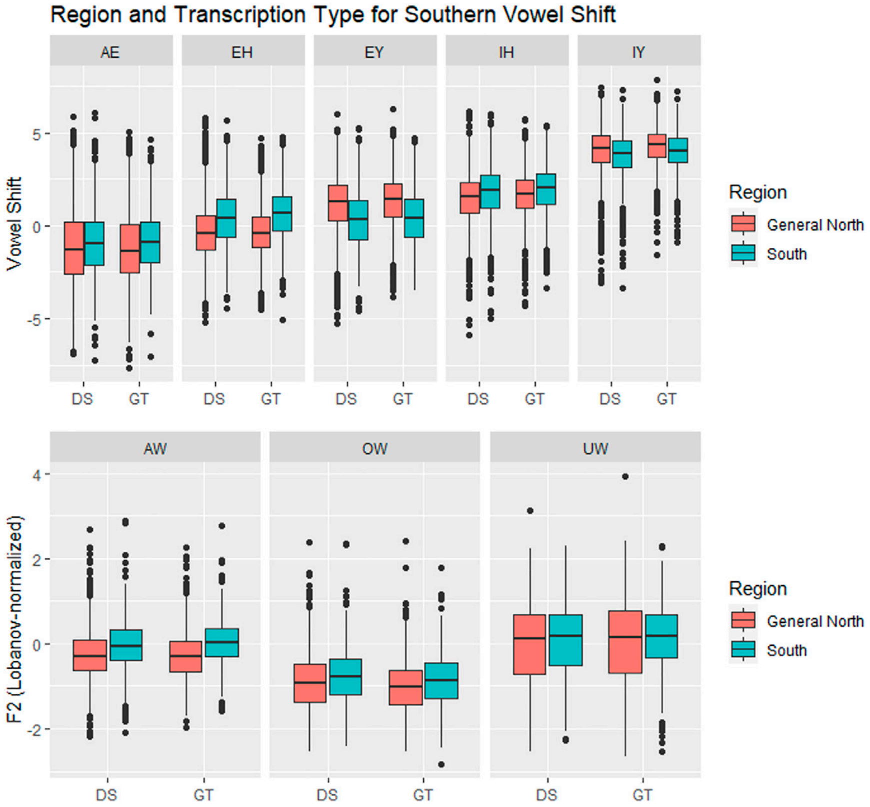


FIGURE 7 | Vowels in the Southern Vowel Shift, by Region (General North versus Southern) and transcription type. In six of the vowels (AW, EH, EY, IH, OW, UW) there is a significant separation between General North and Southern vowels, and this is tracked by both transcription systems. DS = DeepSpeech, GT = Ground truth.

some vowels, like AE, show raising, which is quantified using the standard sociophonetic formula from Labov et al. (2013) 40 which describes such diagonal movement along the front of the vowel space by the relationship: F2 - (2 × F1). Other vowels, such as AW, will use Lobanov-normalized F2 as the dependent variable, to show their movement front or back. All of the dependent variables were transformed (arcsin of the square root) to improve normality and meet the assumptions of LMERs. In summary, the linear mixed-effects modeling will provide 1) a basic description of the Southern Vowel Shift in the data set and 2) a quantified way of determining how close the DeepSpeech transcription gets to the ground truth version. In other words, using publicly available speech recognition methods (like Mozilla DeepSpeech), how close have we come to being able to produce a reliable "hands-free" analysis of a vowel shift from fieldwork recordings of conversations?

Table 2 shows the results from the models. First, we will examine the results related to Region and Transcription type. The most relevant result is that, for seven out of eight vowels involved in the shift, the behavior of the DeepSpeech data is similar to that of the ground truth data. **Figure 7** shows the vowel shift and the F2 for the vowels involved in the Southern Vowel Shift, separated by transcription type. There are five of the vowels, AW, EY, EH, IH, OW, where there are clear differences between the North and South tokens, and these are visible in both the DeepSpeech and the ground truth data. For the vowel AW, for example, the North/South difference for the ground truth is $\Delta\text{Region}_{GT} = 0.31$, whereas the North/South difference for DeepSpeech transcriptions is $\Delta\text{Region}_{DS} = 0.22$. The model as a whole shows differences between North and South [$\beta_{South} = 0.03 \pm 0.01$, t (179) = 2.9, $p < 0.005$]. The model also shows a significant interaction between Region and Transcription [$\beta_{GT:South} = 0.02 \pm 0.007$, t (2567) = 2.2, $p < 0.05$], which means that the $\Delta\text{Region}_{GT} = 0.31$ is significantly smaller than the $\Delta\text{Region}_{DS} = 0.22$. A estimated marginal means (EMM), Tukey-corrected post-hoc analysis was carried out to determine if each of those deltas was actually significantly different from zero (i.e., is there a significant difference between North/South if we looked just at the DeepSpeech data, or if we looked just at the ground truth data?). This was calculated using the emmeans package in R (Russell, 2021). The post-hoc results in **Table 2** confirm that, in the case of AW, both the ground truth ($z = -4.8$, $p < 0.0001$) and DeepSpeech transcriptions ($z = -2.9$, $p < 0.05$) show significant differences between North and South. Taken together, these results indicate that, even if DeepSpeech sees less of a difference between the Northern and Southern tokens of AW, it still sees a significant difference between them, and therefore, the ground truth and DeepSpeech data are describing this sociolinguistic variation is roughly similar ways.

The same general patterns observed for AW are also present in the vowels EY, EH, IH, and OW. In all of them there is a significant difference between Region, which means that the North/South differences were visible in the data. The post-hoc analysis also indicates that the North/South difference can be found in both the ground truth and the DeepSpeech data, and only in the EH is there an interaction between Region and Transcription: The North/South difference in ground truth ($\Delta\text{Region}_{GT} = 0.86$) is significantly larger than the difference in DeepSpeech [$\Delta\text{Region}_{DS} = 0.75$, $\beta_{GT:South} = 0.02 \pm 0.01$, t (144) = 2.0, $p < 0.05$]. In the other three vowels (EY, IH, and OW), both the ground truth and the DeepSpeech data have a similar magnitude for the North/South difference.

There are two vowels, IY and UW, for which neither the ground truth nor the DeepSpeech data could find a significant difference between North and South. (UW has a main effect for Region, but this is an additive effect when both types of transcriptions are put together; once they are separated by the post-hoc test, the significance disappears, with $p = 0.07$ for ΔRegion_{GT} and $p = 0.17$ for ΔRegion_{DS}). This means, in essence, that both ground truth and DeepSpeech data fail to show Northern/Southern differences in similar ways.

The vowel AE deserves special mention because it is the one vowel where ground truth data shows a North/South difference, but DeepSpeech does not. There is a significant difference between transcriptions [$\beta_{GT} = -0.001 \pm 0.0004$, t (265) = −2.6, $p < 0.01$], which is confirmed when the post-hoc results are computed: The ground truth shows a significant difference between North/South AE ($\Delta\text{Region}_{GT} = 0.40$, $z = -3.0$, $p < 0.05$). However, the DeepSpeech data does not show a significant difference in the tokens of AE in the two regions ($\Delta\text{Region}_{DS} = 0.28$, $p = 0.33$). This means that there was one vowel for which DeepSpeech and ground truth disagree. On the other hand, for the other seven, the results from the two transcription types are similar: Either both systems detect a difference, or neither of them does.

The bottom part of **Table 2** also shows results which correspond to patterns that are well established in the sociophonetic literature on the Southern Vowel Shift (Labov et al., 2006). All of the vowels show significant effects for year of birth: For five of the vowels (EY, IY, IH, OW, and UW) younger speakers show more shift, whereas in three of them (AE, AW, and EH), older speakers show more shift. In six of the vowels there are significant differences in shift influenced by the phonological environment of the vowel (e.g., vowel followed by a nasal, a voiced obstruent like/g/, or a voiceless obstruent like/k/). Finally, only one of the vowels (EH) showed a significant difference by gender: Male speakers had a more negative shift (−0.17), whereas female speakers had a more positive shift (0.07).

Table 2 above shows the coefficient of determination (R^2) for the models used. The column shows the marginal correlation coefficient (from the fixed factors) as well as the conditional correlation coefficient (from both the fixed and random factors). The differences between the two show how much of the variation can be explained through random variation due to individual speakers and words: The marginal correlation, which is the correlation from the independent variables, ranges from $R^2 = 0.01$ to $R^2 = 0.21$. On the other hand the conditional correlation, which incorporates the random factor structure, can reach much higher correlation values, up to $R^2 = 0.69$ for the vowel AE, for example. This pattern is to be expected, given that the model doesn't include numerous other factors that could explain variation across speakers (e.g., ethnicity) and variation across words (e.g., lexical frequency). The **Supplementary Appendix** includes the full results for the random variable structure of each model.

In summary, the data from the DeepSpeech automated transcription appears to be adequate in the detection of the SVS vowel patterns. While it is not perfect, it produces similar results to those from manually transcribed data. In the next section we will present evidence of the usability of the DeepSpeech data by focusing on a second sociolinguistic phenomenon, one that has not been extensively studied using automated methods: the vowel shift present in the Northern Cities of the United States.

Northern Cities Shift Results

Figure 8A shows the vowel means of IDEA speakers from Inland North (blue) and the General North regions (red). The General North speakers were defined as all speakers not from Inland North and not from the US Southern regions. In this plot of the results from the ground truth transcription, we see graphical evidence of the five classic NCS vowel movements (Labov et al., 2006; Nesbitt, 2018). First, note that AE has raised for Inland North speakers, representing the classic Stage 1 of the NCS. Second, the Inland North speakers appear to have fronted the AA vowel, which is NCS Stage 2. Then, in the classic chain shift model, the AO vowel has moved toward the original location of the AA vowel, which is Stage 3. For Stage 4, we see that the EH vowel appears to have moved down and back, and for Stage 5, we see that the AH vowel appears to have moved back.

Now consider **Figure 8C**, where we plot the Inland North versus General North again but this time we show the results for DeepSpeech and for the Sphinx ASR transcriptions. Overall, we observe the same NCS shifts in these automated transcription types, with similar directions and magnitudes (e.g., raising of AE). This indicates that the automated methods may be able to uncover the presence of the NCS in these recordings. On the other hand, there are differences between DeepSpeech and Sphinx: Vowels like IY and UW are almost completely overlapping in Sphinx (**Figure 8D**), whereas they show some separation in the DeepSpeech data. In order to test the differences between these, we will first compare Bhattacharyya's Affinity between ground truth/DeepSpeech and ground truth/Sphinx to confirm the improvements from the DeepSpeech transcription. We will then use a linear mixed effects model to confirm that the DeepSpeech data correctly portrays the motions involved in the NCS.

Improvements in Sociophonetic Analysis

We noted in **Figure 8C** that the DeepSpeech system shows graphical evidence of all the same NCS movements as we found above in the ground truth transcription: Raised AE, fronted AA, lowering of AO and EH, and backing of AH. To further the comparison of the transcription methods, all three transcription methods are plotted together in **Figure 9B** below.

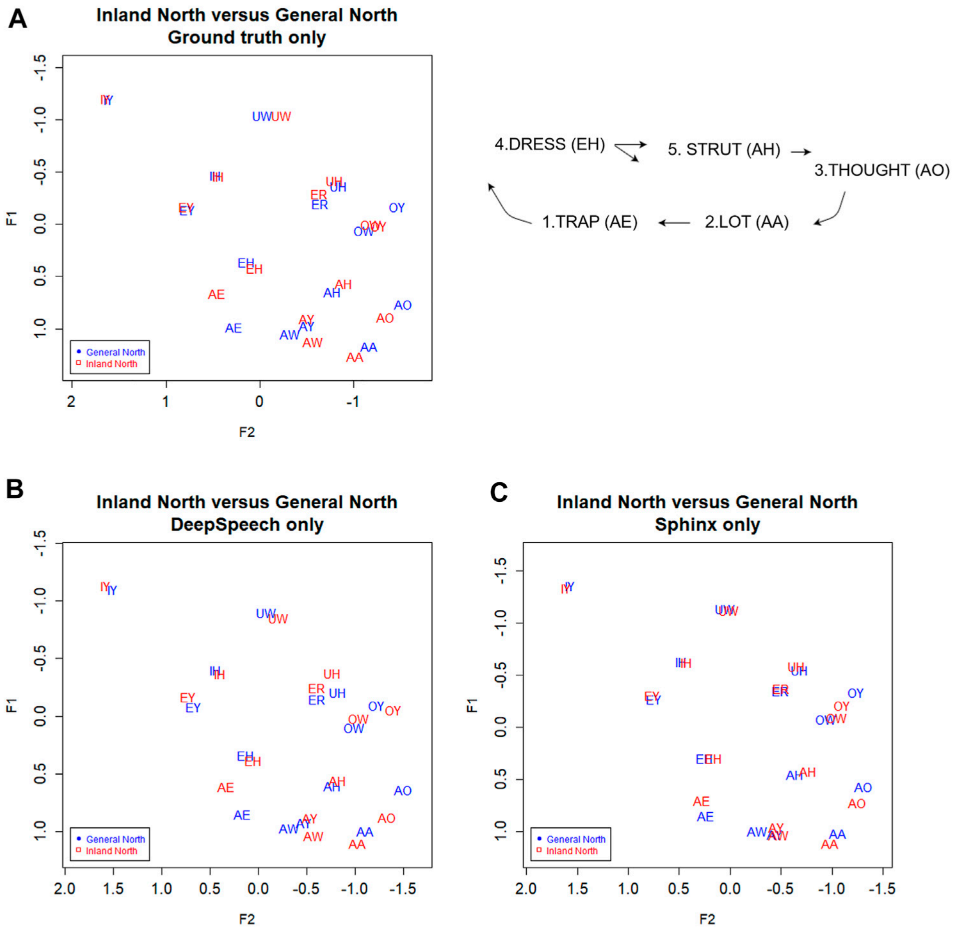


FIGURE 8 | Inland North (red) and General Northern speakers (blue) in the **(A)** ground truth transcription, and transcribed automatically by **(C)** DeepSpeech and **(D)** Sphinx. Speaker vowel means for 225 speakers, 21,200 vowel tokens. Plotted in Lobanov-normalized units. **(B)** Schematic of the primary movements of the Northern Cities Vowel Chain Shift (adapted from Nesbitt, 2018).

(**Figure 4A**, the transcription of the General North vowels, is repeated below as 9a for comparison).

Next, we examine the difference in the NCS vowel transcription statistically. As with the SVS above, we calculated Bhattacharyya's Affinity on Sphinx vs. ground truth and DeepSpeech vs ground truth, and then performed a repeated-measures ANOVA test to determine the effect of transcription and vowels on the BA. We used the same ANOVA structure, with the reflected square root corrected BA as the dependent variable, and vowels and type of transcription as independent, within-subjects variables. There was a significant interaction between vowels and transcription [F (13) = 2.9, $p <$ 0.0005, $\eta^2 = 0.01$] and there is a significant main effect for vowels [F (13) = 9.5, $p <$ 0.0005, $\eta^2 = 0.036$]. This means that there are BA differences between the vowels, and that some vowels benefit more from the DeepSpeech transcription than others. **Figure 10** shows that vowels like AO have a high gain in BA in the DeepSpeech transcription (BA = 0.92 for DeepSpeech but BA = 0.80 for Sphinx). On the other hand, vowels like EY show practically no difference in their Bhattacharyya's affinity, regardless of the transcription mechanism (BA = 0.93 for both DeepSpeech and Sphinx).

There is also a main effect for type of transcription: vowels transcribed with DeepSpeech show higher BAs with the ground truth [F (1) = 29.0, $p <$ 0.00005, $\eta^2 = 0.025$]. In general, the median affinity for DeepSpeech vowels is 0.93, while the median for Sphinx vowels is 0.90. Also, as can be seen in **Table 3**, the vowels involved in the Northern Cities Shift show improvement when transcripted using DeepSpeech. Vowels like EH and AH show only modest differences, whereas the vowel AO shows marked improvement.

Variationist Analysis of NCS Movements

Given the evidence that DeepSpeech is significantly better than CMU Sphinx at transcribing vowels from the Northern Cities Shift, we conducted a linear mixed-effects model analysis of the vowels from the Inland North, comparing DeepSpeech against the baseline ground truth. We use a similar linear mixed-effects structure as in the Southern Vowel Shift above: The fixed variables are Region, Transcription type, Year of birth, Gender, Following environment and the interaction of Region and Transcription type. The random effect structure was also chosen through backward selection using the step procedure;

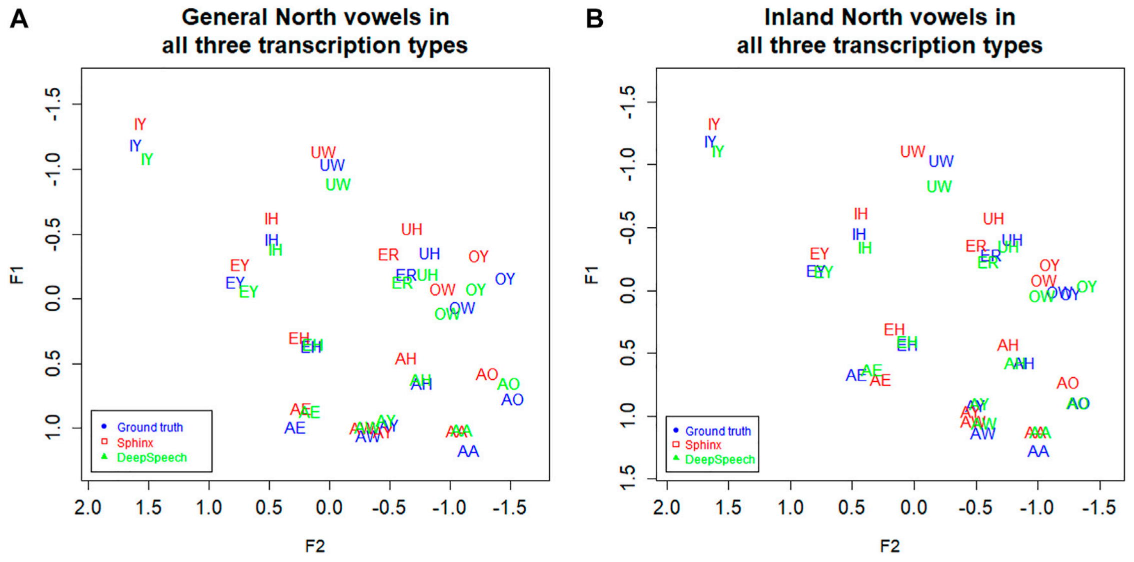


FIGURE 9 | (A) General North and **(B)** Inland North: speaker vowel means in all three transcription types. Blue: Ground truth (manual) transcription, Red: CMU Sphinx transcription Green: DeepSpeech transcription. 58 speakers and a total of 14,414 vowel tokens. Plotted in Lobanov-normalized units.

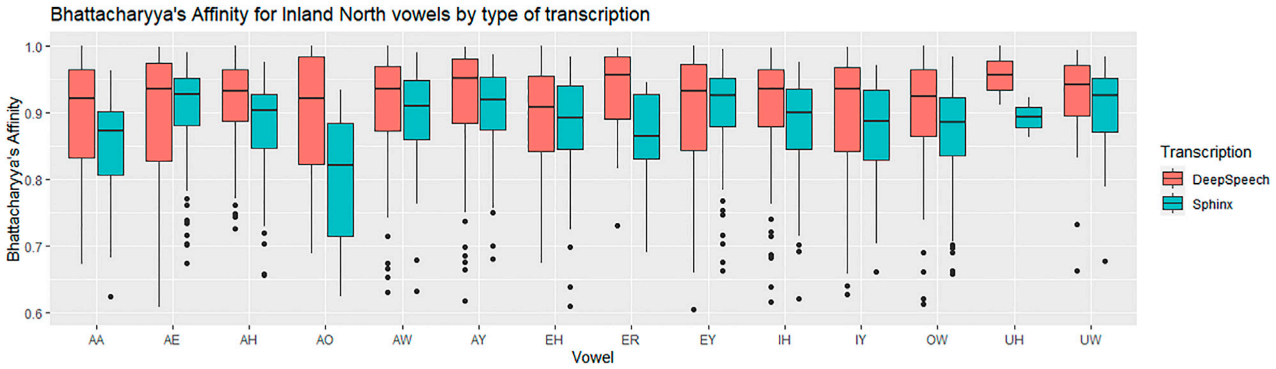


FIGURE 10 | Inland North: vowel medians for the Bhattacharyya's affinity by type of transcription. Red: DeepSpeech versus ground truth, Green = CMU Sphinx versus ground truth. 58 speakers and a total of 14,414 vowel tokens.

the resulting models are shown in the **Supplementary Appendix**.

Following the ANAE (Labov et al., 2006) and Labov (2013:40), we quantify the Northern Cities Shift movements of AA, AH and AO in terms of F2 to characterize fronting. For the diagonal movement of AE along the front of the vowel trapezoid, we follow the method in Labov et al. (2013) 40 of using the equation F2 - (2 × F1) to create a single numerical value representing the raising and fronting. We use the same equation to account for movements of the other front vowel, EH, since the NCS movement of EH may be either backing or lowering or both, as seen above in **Figure 8B**. As with the models above, all of the dependent variables were transformed (arcsin of the square root) to meet the assumptions of linear mixed-effects models.

The results for Region and Transcription type are shown in **Figure 11** and **Table 4A**. For four out of five vowels, the behaviour of DeepSpeech transcribed vowels is similar to that of the vowels in the ground truth transcriptions. There is one vowel, stage 1 AE, where both ground truth and DeepSpeech found significant differences between General and Inland North vowels: ground truth shows a difference of $\Delta\text{Region}_{\text{GT}} = 0.83$ and DeepSpeech shows a significantly smaller but non-zero difference of $\Delta\text{Region}_{\text{GT}} = 0.65$ [$\beta_{\text{GT:InlandNorth}} = -0.02 \pm 0.008$, t (256) = −2.0, $p < 0.05$]. Even though the distance between General and Inland North is smaller for DeepSpeech, it is significantly different from zero, as shown by the post-hoc analysis ($z = 4.8$, $p < 0.0001$). On the other hand, there are three vowels (stage 3 AO and stage 4 EH and stage 5 AH) where neither DeepSpeech nor the ground truth data could see significant differences between General and Inland vowels. For example, EH showed raising differences of $\Delta\text{Region}_{\text{GT}} = 0.21$ and $\Delta\text{Region}_{\text{GT}} = 0.18$, but neither of these were significantly different from zero ($p = 0.88$ and $p = 1.0$ respectively). In summary, for these

TABLE 3 | Bhattacharyya's Affinity for DeepSpeech versus ground truth and Sphinx versus ground truth in Inland Northern speakers. BA score 1.0 = perfectly overlapping distributions, 0.0 = completely non-overlapping.

NCS vowel	Median BA for DS vs. GT	Median BA for SPH vs. GT	Δ(BA)
AA	0.916	0.869	0.047
AE	0.934	0.926	0.007
AH	0.931	0.903	0.028
AO	0.920	0.802	0.118
EH	0.907	0.891	0.016

four vowels (AE, AO, EH, and AH) both systems show similar patterns for both the General and Inland North speakers.

There is one vowel, stage 2 AA, where there is a significant General/Inland North difference in the ground truth data [$\Delta\text{Region}_{GT} = 0.15$, t (94) = 2.9, $p < 0.05$], but there was no significant difference in the DeepSpeech transcriptions ($\Delta\text{Region}_{GT} = 0.07$, $p = 0.76$). This is the one vowel where the two transcription systems diverge. It should be noted that the ground truth data showed differences in the stage 1 and 2 vowels, the ones where the change is presumably more advanced, and it failed to show differences in the subsequent stages 3 through 5. In general, these results show that for most vowels the DeepSpeech data and the manual transcription are similar in how they portray the Northern Cities Shift: Either both of them show the vowel shifts (as is the case for AE) or both of them fail to do so (as is the case for AH, AO, and EH). Only in one of the vowels (AA) was the DeepSpeech data less able to detect the shift.

Like in the case of the Southern Vowel Shift, there are well established Northern Cities Shift linguistic patterns that are visible in the data (Labov et al., 2006). Three out of five vowels show significant differences due to the age of the speakers (stage 1 AE, stage 2 AA, and stage 4 EH): In AE and EH, younger speakers have greater shift; in AA, older speakers have greater shift. Two of the vowels show differences due to gender (stage 1 AE and stage 4 EH): male speakers show greater shift than female speakers. Finally, two of the vowels show differences in the vowel position due to the sounds that follow them (stage 1 AE and stage 4 EH). Also, like in the Southern Vowel Shift data, the R^2 correlation coefficients in **Table 4** show that the random variable structure (individual speaker and word

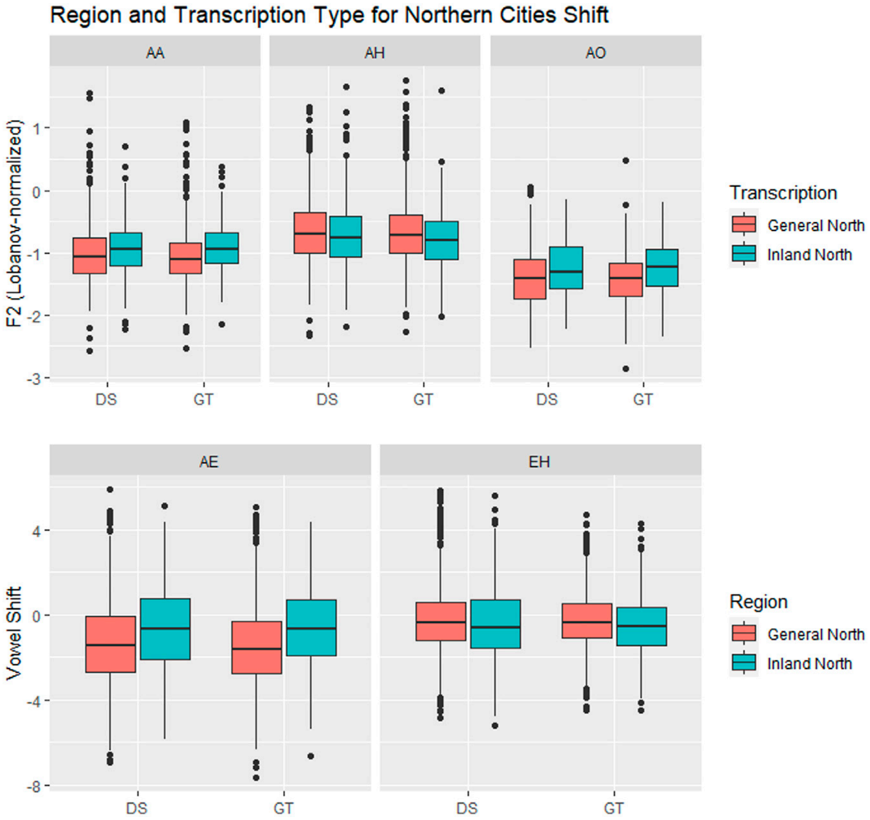


FIGURE 11 | Vowels in the Northern Cities Shift, by Region (General North versus Inland North) and transcription type. In three of the vowels (AE, AA, AO) there are significant differences between General North and Inland North measurements.

TABLE 4 | LMER models for the Northern Cities Shift. The dependent variable is F2 for THOUGHT (AO), LOT (AA) and STRUT (AH), and F2-2xF1 for TRAP (AE) and DRESS (EH). R^2 shows the marginal and conditional coefficients. Deltas show the difference between the mean position of the Southern vowel and the mean position of the Northern vowel for each transcription type: AA shows divergence in GT/DS results; for the other vowels, either both models detect a difference between Inland and General North or they do not.

A. Results for region and transcription type; R^2 for entire model

Vowels	Region by Transcription	Region	Transcription	Post-hoc ΔRegion$_{GT}$	Post-hoc ΔRegion$_{DS}$	R^2
AA (n = 1901)	$p = 0.08$	$p = 0.29$	$p = 0.50$	$\Delta = 0.15$ t (94) = 2.9, $p < 0.05$	$\Delta = 0.07$ $p = 0.76$	0.01 0.72
AE (n = 3433)	$\beta_{GT:InlandNorth} = -0.02 \pm 0.01$ t (256) = −2.0, $p < 0.05$	$\beta_{InlandNorth} = -0.05 \pm 0.01$ t (312) = −4.8, $p < 0.0001$	$p = 0.47$	$\Delta = 0.83$ z = 6.8, $p < 0.0001$	$\Delta = 0.65$ z = 4.8, $p < 0.0001$	0.27 0.67
AH (n = 3388)	$p = 0.13$	$p = 0.24$	$\beta_{GT} = -0.02 \pm 0.007$ t (2703) = −3.0, $p < 0.005$	$\Delta = 0.12$ $p = 0.07$	$\Delta = 0.06$ $p = 0.64$	0.02 0.54
AO (n = 1049)	$p = 0.72$	$p = 0.12$	$p = 0.15$	$\Delta = 0.18$ $p = 0.58$	$\Delta = 0.16$ $p = 0.46$	0.03 0.74
EH (n = 4016)	$p = 0.31$	$p = 0.98$	$p = 0.12$	$\Delta = 0.21$ $p = 0.88$	$\Delta = 0.18$ $p = 1.0$	0.03 0.48

B. Results for other social and linguistic variables in the model

Vowels	Year of birth	Gender	Following environment (Nasal vs. voiced obstruent)	Following environment (Nasal versus voiceless obstruent)
AA (n = 1901)	$\beta = -0.01 \pm 0.004$, t (137) = −2.5, $p < 0.05$	$p = 0.68$	$p = 0.68$	$p = 0.41$
AE (n = 3433)	$\beta = 0.02 \pm 0.004$, t (192) = −3.9, $p < 0.0005$	$\beta_{male} = 0.02 \pm 0.007$, t (191) = 2.5, $p < 0.05$	$\beta_{Nas/+VoicedObs} = -0.12 \pm 0.01$, t (219) = −8.3, $p < 0.00001$	$\beta_{Nas/-VoicedObs} = -0.17 \pm 0.01$, t (359) = −14.9, $p < 0.00001$
AH (n = 3388)	$p = 0.08$	$p = 0.25$	$p = 0.80$	$p = 0.06$
AO (n = 1049)	$p = 0.17$	$p = 0.20$	$p = 0.15$	$p = 0.58$
EH (n = 4016)	$\beta = 0.01 \pm 0.004$, t (172) = −3.8, $p < 0.0005$	$\beta_{male} = 0.02 \pm 0.008$, t (172) = 3.1, $p < 0.005$	$p = 0.42$	$\beta_{Nas/-VoicedObs} = -0.04 \pm 0.01$, t (310) = −3.4, $p < 0.001$

variation) explains significant amounts of the variation found in the dataset: the correlation from the fixed variables ranges from $R^2 = 0.01$ to $R^2 = 0.27$, while the R^2 for the data including the random variable greatly increases (up to $R^2 = 0.74$ in the case of AO). The summary of the variance explained by each of the random variables is in the **Supplementary Appendix**.

In summary, this section provides further evidence that the DeepSpeech data can detect phonetic differences in a manner similar to human-transcribed data. The majority of the vowels involved in the Northern Cities Shift, four out of five, showed a similar behavior in both transcriptions. This matches the pattern we saw with the vowels in the Southern Vowel Shift, where eight out of nine vowels also behaved in a similar manner across both transcription methods.

CONCLUSION

Manual transcription has long been a bottleneck in sociophonetic vowel research. In this paper we have used a large audio dataset of North American English (352 speakers in the International Accents of English Archive) to show that automated speech recognition algorithms (ASR) can be an effective way to perform sociophonetic work for some types of large-scale research questions. We find that end-to-end deep learning based speech recognition algorithms (e.g., DeepSpeech) provide transcriptions that are closer to hand-transcribed data than in prior sociophonetic work. Furthermore, we find that sociophonetic analyses based on these fully automated transcription methods are effective in showing classic sociophonetic patterns of North American English, such as the Southern Vowel Shift (SVS) and the Northern Cities Shift (NCS), with significantly less effort and time invested than manual transcription approaches.

While these DeepSpeech transcriptions are not perfect, we find that they can still be used to gain valuable sociophonetic information. The sociophonetic results derived from the DeepSpeech transcriptions show that the Southern Vowel Shift and the Northern Cities Shift can in fact be graphically observed with these completely automated methods, even as the fine-grained statistical analyses show the ways in which DeepSpeech still lacks the higher degree of precision that can be obtained in analyses based on ground truth (manual) transcription. It also shows that there have been gains in areas relevant to sociolinguistic research, such as the improved transcription of female speech relative to previous ASR methods, as well as the similarity in transcription quality between the standard dialect of North American English and other regional dialects like Southern English.

Much future work remains in order to automatize sociophonetic transcriptions. For example, work needs to be done on whether the method presented here would also detect consonantal sociophonetic variation, given that consonants might not be recognized as reliably as vowels due to their shorter duration. Work also needs to be done on whether this method can be applied to other regional dialects and ethnolects. It is known that English dialects outside of North America, such as New Zealand and Scottish English, are transcribed less accurately (Tatman, 2017), so this method might not be able to detect vowel differences within those dialects. Also, as mentioned above, 79% of the sample was white, and therefore these statistical models might not accurately reflect how the method would perform when transcribing North American ethnolects like Black English, which are not well represented in ASR training corpora (Koenecke et al., 2020). We expect the accuracy of the ASR to be highly variable depending on the types of training input that it received and this could limit the broader application of this method to more diverse datasets. Finally, semi-automated methods like forced alignment have been fruitfully used to phonetics and sociophonetics in languages with extremely small datasets like Yoloxóchitl Mixtec (DiCanio et al., 2013), so there is the potential to apply speech recognition to describe linguistic variation in those languages as well.

Our results suggest that the technology for completely automated methods in vowel sociophonetics is closer to the point where such methods can reliably generate results that are similar to, if not quite the same as, results obtained by the painstaking process of manual transcription. After all, in any scientific endeavor, there is a tradeoff between accuracy and speed, and each research project can determine what type of approach is appropriate. For some sociolinguistic applications and large-scale research questions, such as "big data" analyses of huge sets of audio recordings, it may now be possible to use completely automated methods for reasonably reliable results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.dialectsarchive.com. The code to process the data is available at: https://github.com/rolandocoto/darla-sociophonetics-2021.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RC-S installed Deep Speech in the DARLA system, processed the IDEA recordings through DARLA, conducted quantitative and graphical analyses, contributed computational and statistical knowledge, and wrote the majority of the text of the paper. JS contributed quantitative and graphical analyses, contributed information about U.S. sociophonetics/dialects to the project, and wrote text in the paper. SR designed and built the original DARLA system and contributed her computational knowledge, corrections, and text to the paper.

FUNDING

For this project, we used Dartmouth-internal funding sources to pay for the license to use IDEA recordings and to pay for student research assistants.

ACKNOWLEDGMENTS

The authors would like to thank the editor and the reviewers for their help in improving the paper. We would also like to thank Dagger Bishop, Polina Chesnokova, Catharine Herrera, Rachel Hsu, Cameron Meier, Paul Meier, Veronica Quidore, Andrew Schaeffer, Prof. Monica Williams and Prof. Samantha Wray for their help during the project. Finally, we would like to thank the Dean of Faculty of Dartmouth College for their funding of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.662097/full#supplementary-material

REFERENCES

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., et al. (2019). Common Voice: A Massively-Multilingual Speech Corpus. arXiv preprint arXiv:1912.06670

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal. *J. Mem. Lang.* 68 (3), 255–278. doi:10.1016/j.jml.2012.11.001

Barreda, S. (2021). Perceptual Validation of Vowel Normalization Methods for Variationist Research. *Lang. Variation Change* 33 (1), 27–53.

Barreda, S. (2020). Vowel Normalization as Perceptual Constancy. *Language* 96 (2), 224–254. doi:10.1353/lan.2020.0018

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *J. Stat. Soft.* 67 (1), 1–48. doi:10.18637/jss.v067.i01

Boersma, P., and Weenink, D. (2011). Praat: Doing Phonetics by Computer. Available at: www.praat.org.

Calenge, C. (2006). The Package "adehabitat" for the R Software: A Tool for the Analysis of Space and Habitat Use by Animals. *Ecol. Model.* 197, 516–519. doi:10.1016/j.ecolmodel.2006.03.017

D'Onofrio, A., and Van Hofwegen, J. (2020). "Nisei Style: Vowel Dynamism in a Second-Generation Japanese American Community," in *Speech in the Western States Volume 3*. Editors V. Fridland, A. Wassink, L. Hall-Lew, and T. Kendall (Publication of the American Dialect Society), 105, 79–94.

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., and García, R. C. (2013). Using Automatic Alignment to Analyze Endangered Language Data: Testing the Viability of Untrained Alignment. *The J. Acoust. Soc. America* 134 (3), 2235–2246. doi:10.1121/1.4816491

Evanini, K., Isard, S., and Liberman, M. (2009). Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora. Proceedings of Interspeech. Available at: http://www.isca-speech.org/archive/interspeech_2009/i09_1655.html.

Fridland, V., and Bartlett, K. (2006). The Social and Linguistic Conditioning of Back Vowel Fronting across Ethnic Groups in Memphis, Tennessee. *English Lang. Linguistics* 10 (1), 1–22. doi:10.1017/s1360674305001681

Fridland, V., and Kendall, T. (2019). "5. On the Uniformity of the Low-Back-Merger Shift in the U.S. West and beyond," in *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and Short Front Vowel Shifts across North America*. Editor K. Becker (Publication of the American Dialect Society), 104, 100–119. doi:10.1215/00031283-8032957

Fridland, V., Kendall, T., and Farrington, C. (2014). Durational and Spectral Differences in American English Vowels: Dialect Variation within and across Regions. *J. Acoust. Soc. America* 136, 341–349. doi:10.1121/1.4883599

Godfrey, J., and Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Philadelphia: Linguistic Data Consortium.

Grama, J., and Kennedy, R. (2019). "2. Dimensions of Variance and Contrast in the Low Back Merger and the Low-Back-Merger Shift," in *The Low-Back-Merger Shift: Uniting the Canadian Vowel Shift, the California Vowel Shift, and Short Front Vowel Shifts across North America*. Editor K. Becker (Publication of the American Dialect Society), 104, 31–55. doi:10.1215/00031283-8032924

Hofmann, M. (2014). Mainland Canadian English in Newfoundland. PhD Dissertation. Chemnitz University of Technology.

Hannun, A. Y., Case, Carl., Casper, J., Bryan, C., Diamos, G., Elsen, E., et al. (2014). *Deep Speech: Scaling Upend-To-End Speech Recognition*. ArXiv abs/1412.5567.

Johnson, D. E. (2015). Quantifying Vowel Overlap with Bhattacharyya's Affinity. Presented at NWAV44. Available from: https://danielezrajohnson.shinyapps.io/nwav_44/.

Johnson, D. E. (2009). Getting off the Goldvarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Lang. Linguistics Compass* 3 (1), 359–383. doi:10.1111/j.1749-818x.2008.00108.x

Kendall, T., and Fridland, V. (2021). *Sociophonetics*. Cambridge: Cambridge University Press.

Kendall, T., and Joseph, F. (2014). *Towards Best Practices in Sociophonetics (With Marianna DiPaolo)*. Chicago: New Ways of Analyzing Variation NWAV-43.

Kendall, T., and Thomas, E. (2010). Vowels: Vowel Manipulation, Normalization, and Plotting in R. [R Package]. Available from cran.r-project.org/web/packages/vowels/index.html.

Kendall, T., and Fridland, V. (2012). Variation in Perception and Production of Mid Front Vowels in the U.S. Southern Vowel Shift. *J. Phonetics* 40, 289–306. doi:10.1016/j.wocn.2011.12.002

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2020). Racial Disparities in Automated Speech Recognition. *Proc. Natl. Acad. Sci. USA* 117 (14), 7684–7689. doi:10.1073/pnas.1915768117

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English (ANAE)*. Berlin: Mouton.

Labov, W. (1996). The Organization of Dialect Diversity in North America. *Fourth International* Conference *on Spoken Language Processing*. Available from: https://www.ling.upenn.edu/phono_atlas/ICSLP4.html.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language* 89, 30–65. doi:10.1353/lan.2013.0015

Ladefoged, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Oxford: Blackwell.

Lamere, P., Singh, R., Walker, W., and Wolf, P.Evandro Gouv (2003). "The CMU Sphinx4 Speech Recognition System," in *IEEE Intl. Conf. On Acoustics,Speech and Signal Processing (ICASSP 2003)* (IEEE).

Lenth, R. V. (2021). Emmeans: Estimated Marginal Means, Aka Least-Squares Means. R package version 1.6.0. Available at: https://CRAN.R-project.org/package=emmeans.

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *J. Acoust. Soc. America* 49, 606–608. doi:10.1121/1.1912396

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Morgan, S. (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Proceedings of the 18th Conference of the International Speech Communication Association. doi:10.21437/interspeech.2017-1386

Nesbitt, M. (2018). Economic Change and the Decline of Raised TRAP in Lansing, MI. *Linguistics* 24 (2), 9, 2018 . Available from https://repository.upenn.edu/pwpl/vol24/iss2/9.

Nesbitt, M. (2021). The Rise and Fall of the Northern Cities Shift: Social and Linguistic Reorganization of TRAP in Twentieth Century Lansing, Michigan. *Am. Speech* 96 (3), 332–370.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "Librispeech: an ASR Corpus Based on Public Domain Audio Books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 5206–5210. doi:10.1109/icassp.2015.7178964

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Reddy, S., and Stanford, J. (2015b). "A Web Application for Automated Dialect Analysis," in Proceedings of the North American Association for Computational Linguistics 2015 Conference (NAACL-HLT 2015), 71–75. doi:10.3115/v1/n15-3015

Reddy, S., and Stanford, J. (2015a). DARLA Dartmouth Linguistic Automation: Online Tools for Linguistic Research. Web address: darla.dartmouth.edu.

Reddy, S., and Stanford, J. N. (2015c). Toward Completely Automated Vowel Extraction: Introducing DARLA. *Linguistics Vanguard* 1 (1), 15–28. doi:10.1515/lingvan-2015-0002

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. Available at: http://fave.ling.upenn.edu.

Russell, V. L. (2021). Emmeans: Estimated Marginal Means, Aka Least-Squares Means. version 1.5.3. Available at: https://CRAN.R-project.org/package=emmeans.

Stanley, Joey. (2018). Calculating Vowel Overlap. Available from: https://joeystanley.com/blog/a-tutorial-in-calculating-vowel-overlap.

Strelluf, C. (2016). Overlap Among Back Vowels before/l/in Kansas City. *Lang. Change* 28 (3), 379–407. doi:10.1017/s0954394516000144

Tatman, R. (2017). "Gender and Dialect Bias in YouTube's Automatic Captions," in Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 53–59. doi:10.18653/v1/w17-1606

Tatman, R., and Kasten, C. (2017). "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in Interspeech 2017, 934–938. doi:10.21437/interspeech.2017-1746

Thomas, E., and Kendall, T. (2007). NORM: The Vowel Normalization and Plotting Suite. Online Resource. Available from: http://lingtools.uoregon.edu/norm/norm1_methods.php.

Thomas, E. (2011). *Sociophonetics: An Introduction*. Basingstoke: Palgrave Macmillan.

Warren, P. (2018). Quality and Quantity in New Zealand English Vowel Contrasts. *J. Int. Phonetic Assoc.* 48 (3), 305–330. doi:10.1017/s0025100317000329

Yuan, J., and Liberman, M. (2008). Speaker Identification on the SCOTUS Corpus. *J. Acoust. Soc. America* 123, 3878. doi:10.1121/1.2935783

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Artificial Intelligence

ORIGINAL RESEARCH
published: 01 November 2021
doi: 10.3389/frai.2021.648583

Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter

Quirin Würschinger *

Department of English and American Studies, LMU, Munich, Germany

Societies continually evolve and speakers use new words to talk about innovative products and practices. While most lexical innovations soon fall into disuse, others spread successfully and become part of the lexicon. In this paper, I conduct a longitudinal study of the spread of 99 English neologisms on Twitter to study their degrees and pathways of diffusion. Previous work on lexical innovation has almost exclusively relied on usage frequency for investigating the spread of new words. To get a more differentiated picture of diffusion, I use frequency-based measures to study temporal aspects of diffusion and I use network analyses for a more detailed and accurate investigation of the sociolinguistic dynamics of diffusion. The results show that frequency measures manage to capture diffusion with varying success. Frequency counts can serve as an approximate indicator for overall degrees of diffusion, yet they miss important information about the temporal usage profiles of lexical innovations. The results indicate that neologisms with similar total frequency can exhibit significantly different degrees of diffusion. Analysing differences in their temporal dynamics of use with regard to their age, trends in usage intensity, and volatility contributes to a more accurate account of their diffusion. The results obtained from the social network analysis reveal substantial differences in the social pathways of diffusion. Social diffusion significantly correlates with the frequency and temporal usage profiles of neologisms. However, the network visualisations and metrics identify neologisms whose degrees of social diffusion are more limited than suggested by their overall frequency of use. These include, among others, highly volatile neologisms (e.g., *poppygate*) and political terms (e.g., *alt-left*), whose use almost exclusively goes back to single communities of closely-connected, like-minded individuals. I argue that the inclusion of temporal and social information is of particular importance for the study of lexical innovation since neologisms exhibit high degrees of temporal volatility and social indexicality. More generally, the present approach demonstrates the potential of social network analysis for sociolinguistic research on linguistic innovation, variation, and change.

Keywords: lexicology, lexical innovation, sociolinguistics, diffusion, social media, Twitter, time-series analysis, social network analysis

Edited by:
Jack Grieve,
University of Birmingham,
United Kingdom

Reviewed by:
Alina Maria Cristea,
University of Bucharest, Romania
Aleksei Ioulevitch Nazarov,
Utrecht University, Netherlands

****Correspondence:***
Quirin Würschinger
q.wuerschinger@lmu.de

Specialty section:
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

Received: *31 December 2020*
Accepted: *13 July 2021*
Published: *01 November 2021*

Citation:
Würschinger Q (2021) Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter.
Front. Artif. Intell. 4:648583.
doi: 10.3389/frai.2021.648583

1 INTRODUCTION

Societies continually evolve, new products and practices emerge, and speakers coin and adopt new words when they interact and share information. How do these new words spread in social networks of communicative interaction?

In a recent paper analysing contagion patterns of diseases in *Nature Physics*, Hébert-Dufresne et al. (2020) suggest that the spread of viruses like SARS-CoV-2 follows principles of complex contagion through social reinforcement, and that it matches the dynamics of diffusion of cultural and linguistic innovations such as new words and internet memes. Does this confirm the widespread perception that new words 'go viral'? Influential sociolinguistic models of the spread of linguistic innovations like the S-curve model (Milroy 1992) share fundamental features with earlier economic models of diffusion (Rogers 1962). It is often assumed that diffusion in social networks follows universal trajectories and that rates of spread depend on social dynamics such as network density and the presence or absence of weak ties (Granovetter 1973). Unlike research on biological and cultural diffusion processes, however, sociolinguistic research has only recently been provided with data sources that are equally suitable for large-scale, data-based approaches which can rely on network analyses to study these phenomena empirically.

Social media platforms like Twitter have changed the way we communicate and how information spreads, and they offer valuable data for empirical research. For linguists, social media provides large amounts of data of authentic language use which opens up new opportunities for the empirical study of language variation and change. The size of these datasets as well as their informal nature allow for large-scale studies on the use and spread of new words, for example, to gain insights about general trajectories of diffusion (Nini et al., 2017) or about factors that influence whether new words spread successfully (Grieve, 2018). Moreover, metadata about speakers facilitate the study of aspects of diffusion that go beyond what can be captured by usage frequency alone. Recent work has used Twitter data to investigate the geographical spread of lexical innovations (Eisenstein et al., 2014; Grieve et al., 2016), for example.

Data about the communicative interaction of speakers additionally allows performing network analyses of the social dynamics of diffusion processes. Network science approaches to social media data have been successfully employed in diverse fields, for example, to study the spread of diseases (Lu et al., 2018), opinions (West and Hristo, 2014) and political attitudes (Pew Research Center 2019). While the study of social networks has a long research tradition in sociolinguistics and has shaped influential models of diffusion (e.g., Milroy and Milroy 1985), large-scale network analyses of sociolinguistic phenomena have only recently become more widespread. These new data sources and methodological advances put computational sociolinguistics in an excellent position to gain new insights and to test long-standing theoretical models empirically.

In the area of lexical innovation, this can serve to evaluate important theoretical concepts like the role of early adopters, network density and weak ties in the diffusion of new words. For example, previous approaches have used computational modelling to test the validity of the S-curve model (Blythe and Croft 2012), and to model processes of simple and complex contagion of linguistic innovations in social networks (Goel et al., 2016). Applying social network analysis to bigger samples of neologisms and tracking their use and spread on social media datasets promises to provide a more detailed picture of social diffusion. Social network information has the potential to more accurately assess the degrees to which the adoption of new words remains limited to closely connected sub-communities or whether they reach larger parts of the speech community.

This paper aims to explore the role of network information and temporal dynamics in assessing the diffusion of lexical innovations on Twitter. I use several quantitative and qualitative methods to study diffusion. I conduct a longitudinal study monitoring the use of a broad sample of neologisms to analyse their usage frequency and the temporal dynamics underlying their use. Next, I use social network analyses to get a better picture of the sociolinguistic dynamics at play, to assess different pathways and overall degrees of diffusion. Lastly, I combine both approaches to get a more detailed picture of the diffusion of the neologisms in the sample, and to assess the results of both approaches to diffusion.

The paper is structured as follows. **Section 2** introduces the theoretical framework for modelling and measuring the diffusion of lexical innovations which forms the basis for the empirical study. **Section 3** presents information about the sample of neologisms and the Twitter dataset this study is based on. **Section 4** describes the methods used for analysing diffusion. **Section 5** presents the results of the empirical study. I analyse diffusion on the basis of frequency and social networks and integrate the results obtained from both approaches. **Section 6** summarises and discusses the results from the empirical study and draws implications about the role of frequency and network-based measures for the study of diffusion.

2 MODELLING AND MEASURING THE DIFFUSION OF LEXICAL INNOVATIONS

2.1 Modelling Diffusion

Neologisms are on a continuum from entirely novel word-formations to fully established lexemes which are familiar to the majority of the speech community. Neologisms have spread to some extent, but are still perceived as new or unknown by many speakers (Schmid 2016). On one end of the continuum, 'ad-hoc formations' are new words that have been coined in a concrete communicative situation, but are not adopted by interlocutors and do not diffuse beyond their original usage contexts (Hohenhaus 1996). On the other end, fully established words are known and used by the majority of the speech community. Neologisms occupy an intermediate position between both poles and can be defined as '(. . .) lexical units, that have been manifested in use and thus are no longer nonce-formations, but have not yet occurred frequently and are not widespread enough in a given period to have become part and parcel of the lexicon of the speech community and the majority of its members' (Kerremans 2015, 31).

Diffusion can be seen as the process that transports successful neologisms along this continuum while they are becoming increasingly conventional in the speech community. The S-curve model (Milroy 1992; Nevalainen 2015; Labov 2007) expects an S-shaped trajectory for the spread of linguistic innovations and makes specific assumptions about the sociolinguistic characteristics of speakers involved in the diffusion process. In a first stage of slow diffusion, only a small number of early adopters take up the innovative words. These individuals typically form dense networks which are connected by strong ties. In the case of successful diffusion, the initial stages are followed by an acceleration in spread when new words increasingly reach speakers outside the initial communities. Weak ties (Granovetter 1973) play an important role in allowing the innovations to reach a bigger parts of the speech community. During later stages, rates of diffusion slow down again as the majority of the speech community has already adopted the new words, while a minority of speakers remains resistant to take up the new words.

The Entrenchment-and-Conventionalization Model (Schmid 2020) conceptualises the conventionalization of linguistic innovations as involving two processes: usualization and diffusion. Diffusion is defined as the process that 'brings about *a change in the number of speakers and communities* who conform to a regularity of co-semiotic behaviour and a change in the conformity regarding the types of cotexts and contexts in which they use it.' (Schmid 2020, 178–179, emphasis mine) In the case of a given new word, it is coined by an individual speaker and first reaches a community of speakers who might be closely-connected to the coiner and/or share interests related to the given neologism. With more advanced diffusion, the word spreads to larger numbers of speakers and increasingly also becomes conventional in other communities of speakers. The process of usualization, by contrast, leads to the increasing establishment of a given neologism by repeated use within one community of speakers. Neologisms thus show high degrees of conventionality, when they exhibit high usage intensity across a large number of speakers and communities.

2.2 Measuring Diffusion

Earlier empirical work on lexical innovation had to rely on smaller, general-purpose linguistic corpora. The low-frequency nature of neologisms limited earlier studies to conducting case studies on selected neologisms (Hohenhaus 1996) or on specific domains of neology (Elsen 2004). In recent years, research on lexical innovations has seen an upsurge in large-scale empirical investigations on the diffusion of neologisms, thanks to the availability of new data sources and computational methods.

The increasing availability of web corpora significantly extended the opportunities for large-scale corpus analyses. Modern corpora like the NOW corpus (Davies 2013) allow to study more comprehensive samples of neologisms and enable researchers to monitor their use over time, which is essential for investigating diffusion processes. In addition to general-purpose web corpora, several research groups built dedicated tools and specialized corpora for the monitoring and analysis of neologisms (Renouf et al., 2007; Kerremans et al., 2012; Lemnitzer, 2010; Gérard et al., 2017; Cartier 2017).

More recently, social media data have become an increasingly important alternative to web corpora. Language use on social media is informal and creative, which makes it a hotbed for lexical innovation. Recent work using Twitter data has focused, for example, on the identification of neologisms (Grieve et al., 2018), on their geographical diffusion (Eisenstein et al., 2014), and on trajectories of diffusion (Nini et al., 2017). Empirical investigations on the basis of Reddit data include studies of the linguistic dissemination of neologisms (Stewart and Jacob. 2018) and the role of innovators and adopters (Del Tredici et al., 2018).

The present study is based on Twitter data and goes beyond previous work in its focus on the sociolinguistic dynamics of diffusion, which are at the core of theoretical models of diffusion. Most previous empirical investigations of the spread of new words have been limited to using frequency measures as an indicator of diffusion. While frequency counts have proven useful in previous work, they can only provide limited insight into the sociolinguistic dynamics of diffusion (Stefanowitsch and Flach 2017). In addition to usage frequency, I will therefore use network information to assess the social pathways of diffusion in the present dataset.

3 DATA

3.1 Neologism Sample

The present study is based on a selection of 99 neologisms and investigates their use on Twitter from its launch in 2006 to the end of 2018. The lexemes were selected to cover a broad spectrum of lexical innovation. Previous work by Kerremans (2015, 115–147) has identified four main clusters of neologisms on the conventionalization continuum: 'non-conventionalization', 'topicality or transitional conventionalization', 'recurrent semi-conventionalization' and 'advanced conventionalization'. The present sample was designed to cover these categories and largely contains neologisms taken from the NeoCrawler (Kerremans et al., 2012), which uses dictionary-matching to retrieve a semi-automatic, bottom-up selection of recent neologisms on the web and on Twitter (Kerremans et al., 2019). I have additionally included several lexemes that were statistically identified to have been increasing in frequency on Twitter in recent years by Grieve et al. (2016). I limit my selection to neologisms whose diffusion started after 2006 to have full coverage of the incipient stages of their spread on Twitter.

3.2 Twitter Corpus

Twitter is a popular micro-blogging platform that was started in 2006 and has become one of the most popular social media platforms today. Its broad user base and informal nature allow for a more representative picture of language use than domain-specific studies of, for example, newspaper corpora.[1] Twitter corpora have been successfully used to identify patterns of sociolinguistic variation in numerous previous studies. A

[1] The present dataset was restricted to tweets in the English language. Due to the absence of the required metadata, the data cannot be further restricted to specific geographical regions, and it is not possible to identify native speakers of English.

recent study by Grieve et al. (2019), for example, has demonstrated the reliability of large-scale Twitter datasets for studying lexical variation.

Twitter is particularly well-suited for studying lexical innovation due to the scale and types of data it provides, and due to the nature of language use on Twitter. The large size of Twitter's search index facilitates the quantitative study of neologisms, which requires large-scale datasets due to their inherently low frequency of occurrence. Twitter is widely used to discuss trends in society and technology, which makes it a good environment for studying the emergence of linguistic innovations. The informal and interactional nature of communication on Twitter fosters the rapid adoption of linguistic innovations, and the use of neologisms on social media platforms like Twitter often precedes and drives the diffusion of new words in more formal sources or on the web (Würschinger et al., 2016).

The data for this study were collected using the Python library *twint*, which emulates Twitter's Advanced Search Function. For each word in the sample, I performed a search query to retrieve all tweets found in Twitter's search index. Due to the large volume of more frequent lexemes, I limited the sample to contain only candidates for which I could collect all entries found in Twitter's index. The combined dataset for all 99 lexemes in the sample contains 29,912,050 tweets. The first tweet dates from May 5, 2006 and involves the neologism *tweeter*, the last tweet in the collection is from December 31, 2018, and includes *dotard*.

4 METHODS

I processed the dataset to remove duplicates, tweets that do not contain tokens of the target neologism in the tweets' text body. This was mostly relevant in cases where Twitter returned tweets in which the target forms were only part of usernames or URLs.[2] Hashtag uses were included in the analysis. Retweets were excluded, since the data did not provide reliable information about retweeting activity for the social network analysis. The resulting dataset contains about 30 million tweets, and each tweet contains at least one instance of the 99 neologism under investigation.

To investigate the diffusion of these lexemes in terms of usage frequency, I use time-series of the neologisms' frequency of occurrence over time. I binned the number of tweets per lexeme in monthly intervals to weaken uninterpretable effects of daily fluctuations in use, and to achieve a reasonable resolution to compare the use of all lexemes, which differ according to their overall lifespan. I visualize the resulting time series as presented in **Figure 2**.

To capture different degrees of stability vs. volatility in the use of neologisms over time, I calculated the coefficient of variance for all time series. The coefficient of variance (c_v) is a measure of the ratio of the standard deviation to the mean: $c_v = \frac{\sigma}{\mu}$. Higher values indicate higher degrees of variation in the use of a neologism, which is typical of topical use of words such as *burquini*; lower values indicate relatively stable use of words such as *twitterverse*.

To investigate the diffusion across social networks over time, I subset the time series into four time frames of equal size, relative to the total period of diffusion observed for each neologism. I set the starting point of diffusion to the first week in which there were more than two interactions which featured the target lexeme. This threshold was introduced to distinguish early, isolated ad-hoc uses of neologisms by single speakers from the start of accommodation processes during which new words increasingly spread in social networks of users on Twitter. This specific limit was determined and validated empirically by systematically testing different combinations of threshold values for the offset of number of users and interactions among early users. Setting a low minimum level of interactions per week proved to reduce distortions in the size of time windows, and enabled a more robust coverage of the relevant periods of diffusion. For each neologism, I divided the time window from the start of its diffusion to the end of the period covered by the dataset into four equal time slices that are relative to the varying starting points of diffusion for all words in the sample. The starting points of each time frame are marked by dashed vertical lines in the usage frequency plots presented below (**Figure 2**).

To investigate the social dynamics of diffusion over time, I generated social networks graphs for each of these subsets. Nodes in the network represent speakers who have actively used the term in a tweet and speakers who have been involved in usage events in the form of a reply or a mention in interaction with others. The resulting graphs represent networks of communicative interaction. Communities are formed based on the dynamic communicative behaviour observed, rather than on information about users' social relations as found in follower–followee networks. This methodology is supported by previous research, which suggests that interactional networks of this kind are better indicators of social structure, since the dynamic communicative behaviour observed is more reliable and socially meaningful than static network information (Goel et al., 2016; Huberman et al., 2008). While users often follow thousands of accounts, their number of interactions with others provides a better picture of their individual social networks, which are much more limited in size (Dunbar 1992).

To construct the networks, I extracted users and interactions from the dataset to build a directed graph.[3] Nodes in the graph correspond to individual Twitter users, edges represent interactions between users. I captured multiple interactions between speakers by using edge weights, and I accounted for active vs. passive roles in interaction by using directed edges. I assessed the social diffusion of all neologisms quantitatively by generating and comparing several network metrics, and I

[2] The post-processing and all quantitative analyses were performed in R Core Team (2018), and the source code is available on GitHub: https://github.com/wuqui/sna.

[3] I used several *R* packages (R Core Team 2018) from the *tidyverse* library collection (Wickham et al., 2019) for the network pre-processing; *igraph* and *tidygraph* were used for constructing the networks and for calculating network metrics.

produced network visualisations for all subsets for more detailed, qualitative analyses.

On the graph level, I rely on the measures of *degree centralization* and *modularity* to quantify the degree of diffusion for each subset. Degree centralization (Freeman 1978) is a graph-level measure for the distribution of node centralities in a graph. Nodes have high centrality scores when they are involved in many interactions in the network and thus play a 'central' role in the social graph of users. The degree centrality of a graph indicates the extent of the variation of degree centralities of nodes in the graph. A graph is highly centralized when the connections of nodes in the network are skewed, so that they center around one or few individual nodes. In the context of diffusion, the graph of a neologism tends to have high centralization in early stages when its use is largely confined to one or few centralized clusters of speakers. Diffusion leads to decreasing centralization when use of the term extends to new speakers and communities and the distribution of interactions in the speech community shows greater dispersion.

The normalized degree centralization of a graph is calculated by dividing its centrality score by the maximum theoretical score for a graph with the same number of nodes. This enables the comparison of graphs of different sizes, which is essential for drawing comparisons across lexemes in the present context. The neologisms under investigation differ with regard to their lifespan and usage intensity, resulting in substantial quantitative differences in network size. This needs to be controlled for to allow for an investigation of structural differences of the communities involved in their use.

Modularity (Blondel et al., 2008) is a popular measure for detecting the community structure of graphs. It is commonly used to identify clusters in a network and provides an overall measure for the strength of division of a network into modules. In the social context, this corresponds to the extent to which the social network of a community is fragmented into sub-communities. Networks with high modularity are characterized by dense connections within sub-communities, but sparse connections across sub-communities. In the context of the spread of new words on Twitter, diffusion leads from use limited to one or few densely connected communities to use in more and more independent communities. This is reflected by higher degrees of modularity of the full graph representing the speech community as a whole. Modularity complements degree centralization since it provides additional information about the number and size of sub-communities who use the target words. I rely on the modularity algorithm to perform community detection, and I visualize the eight biggest communities in each graph by colour.

Since modularity is sensitive to the number of edges and nodes in a graph and thus cannot provide reliable results for comparing graphs of different size, I use degree centralization to analyse diffusion over time, and to assess differences in degrees of diffusion between lexemes on the macro-level. Its conceptual clarity and reliable normalization allow for more robust comparisons on the macro-level.

For visualizing network graphs, I rely on the Force Atlas 2 algorithm (Jacomy et al., 2014) as implemented in *Gephi* (Bastian et al., 2009). Force Atlas 2 is a force-directed algorithm that attempts to position the graph's nodes on a two-dimensional space such that edges should be of similar length and there should be as little overlap between edges as possible. In the present social network graphs, the algorithm places nodes (speakers) closer to each other if they have one or more edges connecting them (communicative interactions in the form of replies and mentions). Attempts to evaluate and compare these visualisations with results obtained from different algorithms such as Multi-Dimensional Scaling and Kamada Kawai showed similar results across methods for parts of the dataset, but could not be used for the full dataset due to the computational complexity involved in the generation of large-size graphs of high-frequency neologisms. Force Atlas 2 is particularly well-suited for handling social networks in big data contexts and has been widely applied in network science approaches to Twitter data (Bruns 2012; Bliss et al., 2012; Gerlitz and Rieder 2013).

To assess and visualize the influence of individual users in the social network, I use the PageRank algorithm (Brin and Page 1998). PageRank assesses the importance of nodes in a network based on how many incoming connections they have. It was initially used to analyse the importance of websites on the World Wide Web, but it is also frequently applied to determine the influence of agents in social networks (e.g., Halu et al., 2013; Pedroche et al., 2013; Wang et al., 2013). In the present context, PageRank assigns higher scores to speakers who receive more incoming replies and mentions, which I visualise by bigger node sizes in the network graphs. To account for varying degrees of strength in the connection between users, I use edge weights for repeated interactions, visualised by the edges' width in the graphs.

5 RESULTS

5.1 Frequency-Based Measures of Diffusion

5.1.1 Overall Usage Frequency

As described in **Section 2.1**, successful diffusion involves an increase in the number of speakers and communities who know and use a new word. The degree of diffusion of new words is often approximated by usage frequency, i.e., by how many times speakers have used a given word in the corpus. The most fundamental way of using this information is to aggregate usage counts and to rely on the total number of uses observed. The underlying assumption is that neologisms that have been used very frequently in the corpus are likely to be familiar to a large group of speakers who have actively produced the observed uses ('corpus-as-output') or have been passively exposed to these neologisms ('corpus-as-input') (Stefanowitsch and Flach 2017). Aggregating all instances of usage to total counts is taken to represent the total amount of exposure or active usage, indicating the degree of conventionality in the speech community. In the following, I will use this most basic measure of diffusion as a baseline before I zoom in to get a more differentiated picture of the temporal and social dynamics of diffusion.

The present sample of neologisms covers a broad spectrum of usage frequency. **Tables 1–4** presents the candidates under investigation in four groups: six examples around the

TABLE 1 | Total usage frequency (FREQ) in the corpus. Most frequent lexemes.

Lexeme	FREQ
tweeter	7,367,174
fleek	3,412,807
bromance	2,662,767
twitterverse	1,486,873
blockchain	1,444,300
smartwatch	1,106,906

TABLE 2 | Total usage frequency (FREQ) in the corpus. Examples around the median.

Lexeme	FREQ
white fragility	26,688
monthiversary	23,607
helicopter parenting	26,393
deepfake	20,101
newsjacking	20,930
twittosphere	20,035

TABLE 3 | Total usage frequency (FREQ) in the corpus. Least frequent lexemes.

Lexeme	FREQ
microflat	426
dogfishing	399
begpacker	283
halfalogue	245
rapugee	182
bediquette	164

TABLE 4 | Total usage frequency (FREQ) in the corpus. Case study selection.

Lexeme	FREQ
alt-right	1,012,150
solopreneur	282,026
hyperlocal	209,937
alt-left	167,124
upskill	57,941
poppygate	3,807

minimum, around the median, and around the maximum total usage frequency observed in the corpus, as well as six words that will serve as case studies in the following sections. These cases reflect a set of prototypical examples of different pathways of diffusion, and I will use these cases to illustrate more detailed characteristics of diffusion before I present the general patterns found for the full sample of neologisms.

The grouping of neologisms on the basis of their total usage frequency presented in **Tables 1–4** largely seems to fit intuitions about diverging degrees of conventionality between the frequency-based groups listed in **Tables 1–3**. Neologisms such as *blockchain* and *smartwatch*, which are probably familiar to most readers, can be assumed to be more conventional than neologisms from the low end of the frequency continuum such as *dogfishing* ('using a dog to get a date') or *begpacker* ('backpackers funding their holidays by begging').

However, total frequency counts only provide a limited picture of diffusion since they are insensitive to temporal dynamics of usage. Neglecting temporal information about the lifespan and the period of active use of a new word can distort the quantitative assessment of its degree of conventionality in two directions. Firstly, it carries the danger of overestimating the status of words such as *millennium bug*[4], whose total usage frequency largely goes back to a short period of highly intensive usage, after which they fall into disuse, become unfamiliar to following generations of speakers, eventually becoming obsolete. Secondly, total counts can underestimate the conventionality of words such as *coronavirus*, which have already become familiar to the vast majority of speakers, but show comparatively moderate total frequency counts, since they have started to diffuse only fairly recently.

Among the most frequent neologisms presented in **Table 1**, words such as *twitterverse* and *blockchain*, for example, have similar total frequency counts, but differ significantly with regard to their temporal usage profiles. The neologism *twitterverse* has been in use ever since the start of Twitter, while the diffusion of the much younger *blockchain* only started in 2012. Despite its shorter lifespan, *blockchain* accumulated roughly the same number of uses, but shows significantly higher usage intensity in the more recent past, and can be assumed to be familiar to bigger parts of the speech community.

Similar effects are even more pronounced in the remaining groups of neologisms, since words from the lower ranges of the frequency spectrum are typically affected more strongly by temporal variation in their use. In the following sections, I will include temporal information to get a more fine-grained picture of diffusion.

5.1.2 Cumulative Frequency

Visualising the cumulative increase in usage frequency of new words complements total counts by taking into account the temporal dynamics of their usage intensity over time. **Figure 1** presents this information for the case study selection.

While the end points of the trajectories in **Figure 1** mark the target words' total frequency counts as shown in **Table 4**, the offsets and slopes of the trajectories of usage frequency reveal additional characteristics about differences in their diffusion patterns. The selected neologisms differ regarding their total lifespan observed, which is indicated by diverging starting points of diffusion. The term *hyperlocal*, for example, is the oldest new word among the selected neologisms, and it is commonly used to refer to information that has a strong focus on local facts and events. While it was hardly used in the first years of Twitter, it started to increase in its use in 2009 and was added to the OED's Third Edition in 2015. Around this time, the neologism *solopreneur* only started to significantly increase in its use. A blend of *solo* and *entrepeneur*, it keeps a low, flat trajectory

[4] The neologisms *millennium bug* was used to refer to ancipated technical problems caused by inconsistent formatting of timestamps at the turn of the century.

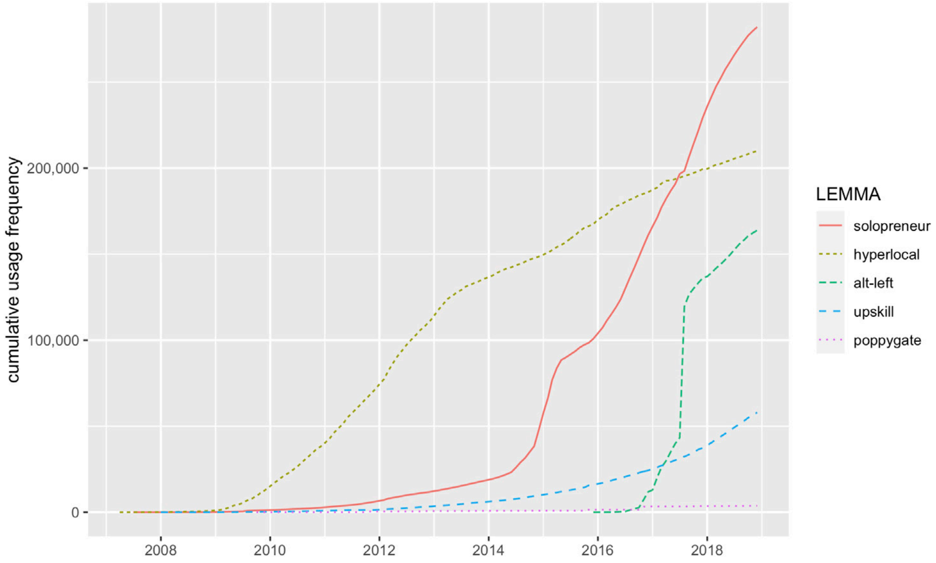


FIGURE 1 | Cumulative increase in usage frequency for the case study lexemes[5].

of sporadic use for about 7 years after its first appearance in the corpus. The first two attestations in the corpus indicate the sense of novelty and scepticism towards the term in its early phases:

1) I'm trying to figure out if I like the term 'solopreneur' I just read (July 27, 2007).
2) hmmmmmmm new word added to my vocab = 'solopreneur' !! (January 6, 2008).

Most speakers increasingly 'like the term' and 'add them to their vocabulary' only much later, after 2014, when the phenomenon of individual entrepreneurship attracts increasing conceptual salience in the community, which seems to be both reflected and propagated by the publication of several self-help books for entrepreneurs in this year, which all explicitly use this new term in their titles (e.g., the popular guide *Free Tools for Writers, Bloggers and Solopreneurs* by Banes (2014)). The following short, but intense period of use results in a higher overall number of uses for *solopreneur* as compared with *hyperlocal*, even though the use of the latter term shows a longer lifespan of continual use[5].

In addition to differences in age, the slopes of the cumulative trajectories in **Figure 1** indicate differences regarding the dynamics of diffusion underlying the aggregated total number of uses over time.

Neologisms such as *hyperlocal* and *upskill* ('to learn new skills') show a steady, gradual increase in usage frequency over longer periods of time. By contrast, the use of other candidates such as *solopreneur* and *alt-left* is much less stable and less evenly distributed over time.

In the case of *solopreneur*, we observe a big spike in frequency following its increased popularity in the entrepreneurial community in 2014. While it shows the highest total frequency count in **Figure 1**, the majority of its uses fall into the second part of its observed lifespan.

An even shorter and steeper increase can be seen in the use of *alt-left*, which is the youngest neologism to enter the scene at the end of 2015. *alt-left* was coined as a counterpart to the term *alt-right*. The latter neologism is a shortening of *Alternative Right*, introduced by the white-supremacist Richard Spencer in 2010 as a new umbrella term for far-right, white nationalist groups in the United States. Facing substantial criticism for racist attitudes and actions, proponents of this far-right political camp coined and attempted to propagate the derogatory term *alt-left* to disparage political opponents. Despite its late appearance in the corpus, *alt-left* occurs in a total of 163,809 tweets, which places it in the medium range of the sample in terms of total frequency counts. However, its trajectory in **Figure 2** shows that the majority of its uses go back to a single period of highly intensive use in the second half of 2017, soon after which it slows down considerably.

The cumulative increase in usage intensity of the selected neologisms illustrates that similar total frequency counts of neologisms can be the product of highly different trajectories of diffusion. These data complement total counts in that they show differences in the total lifespan and in the intensity with which a given neologism was used over time – types of information that are highly relevant for assessing the degree to which they have spread in the speech community.

5.1.3 Usage Intensity

Going beyond cumulative counts, absolute usage frequency counts provide a more fine-grained view of the temporal dynamics of diffusion. Most importantly, analysing usage

[5] *alt-right* was omitted from this plot because its high usage frequency would have inhibited the interpretability of the other lexemes; its frequency over time is presented in **Figure 3D**.

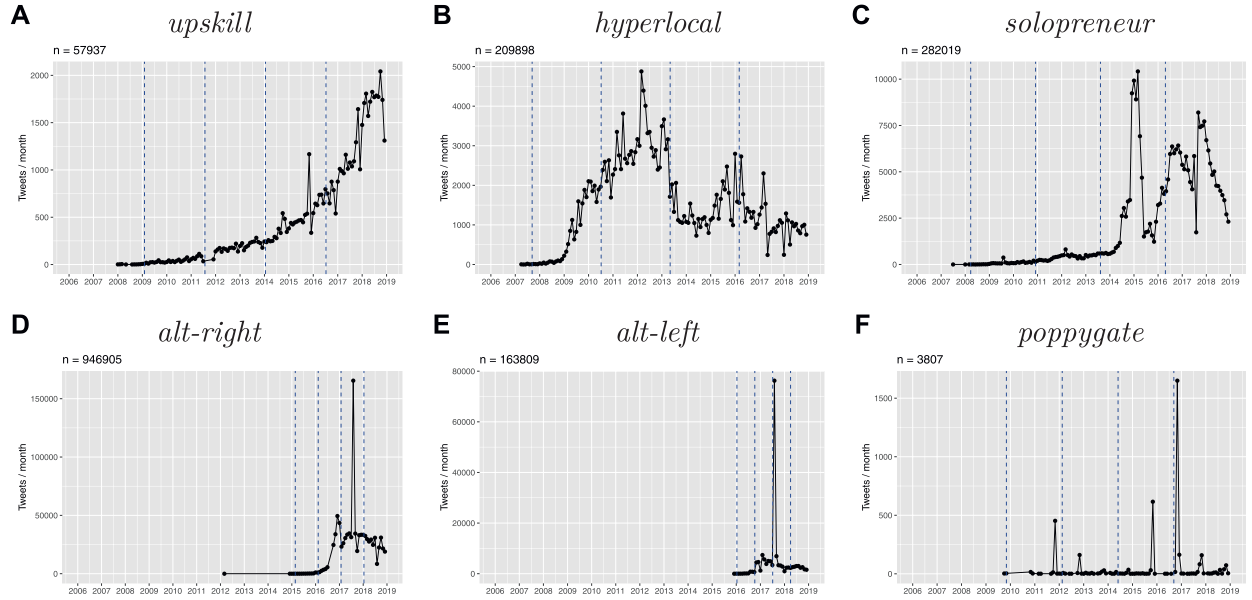


FIGURE 2 | Temporal dynamics in usage frequency for the selected neologisms.

intensity highlights to what degree new words are being used consistently over time. **Figure 2** presents this information for the selected neologisms. In the following section, I will illustrate prototypical differences by referring to the selected cases, before I discuss the results for the full sample[6].

The absolute frequency plots confirm differences regarding the lifespan and dynamics of usage intensity among the neologisms discussed above. In terms of lifespan, **Figure 2** shows that *upskill* and *hyperlocal* are much older than *alt-right* and *alt-left*. The absolute counts also highlight the fact that while there is a low level of use of *solopreneur* since 2007, its main period of diffusion starts much later, in 2014, with a subsequent spike in usage intensity.

5.1.4 Volatility

Besides, the absolute frequency counts over time provide a more detailed picture of the temporal dynamics of use. While the cumulative counts in **Figure 1** suggest more gradual trajectories, the plots in **Figure 2** indicate that the selected neologisms differ significantly in terms of the volatility with which they are used in the corpus.

The neologism *upskill* shows the smoothest trajectory of diffusion among the candidate neologisms in **Figure 2**. Aside from two smaller spikes, at the end of 2016 and 2018, it has gradually increased in its use since its first attestation in the corpus at the end of 2007. Neither its frequency counts, nor the corpus data suggest that its spread was triggered or propagated by specific topical events or by the determining influence of individual users or user groups. After a long period of very slow, but consistent increase in frequency, its diffusion has accelerated in recent years. While its future remains uncertain, its previous trajectory resembles most closely the earlier phases of spread as predicted by S-curve models.

While *hyperlocal* also exhibits a marked increase in usage frequency during its earlier stages, its peak in popularity is followed by a decline in use, after which it settles at a relatively stable level of about 1,000 tweets per month. This coincides with the OED's decision to take up *hyperlocal* in its 2015 edition. Despite fluctuations, *hyperlocal* has been used relatively consistently in the recent past.

The neologism *solopreneur* has been in use since 2007 and shows an overall increase in usage frequency, but its use fluctuates more strongly than that of *hyperlocal*. After its initial peak around 2015, which coincides with the release of several self-help books featuring the term, its frequency plummets, becomes less stable, and shows an overall downward trend.

As was mentioned above, *alt-right* and *alt-left* are closely related. Both terms show high levels of volatility in their usage frequency. The former, older term shows significant diffusion in 2016, particularly in the period leading up to Donald Trump's election, after which *alt-right* remains in consistent use to a relatively high degree, at about 25,000 tweets per month. Its counterpart, *alt-left*, enters the scene much later, during the infamous Charlottesville Rally in 2017, whose topical effect causes a huge spike in the use of both terms. However, unlike *alt-right*, which reverts to its previous usage intensity, the use of *alt-left* seems to largely disappear from Twitter in the aftermath of the event.

The final example among the selected candidates, *poppygate*, also exhibits high degrees of volatility, and it features the most distinctive pattern of spikes in its usage intensity. Unlike the single topical spike for *alt-right* and *alt-left*, its use follows a recurrent, regular pattern: speakers use it almost exclusively

[6] Neologisms with a lifespan shorter than 1 year and/or less than 2,000 tweets (n = 5) were excluded since the coefficient of variation does not provide robust measures for these infrequent, short-lived outliers.

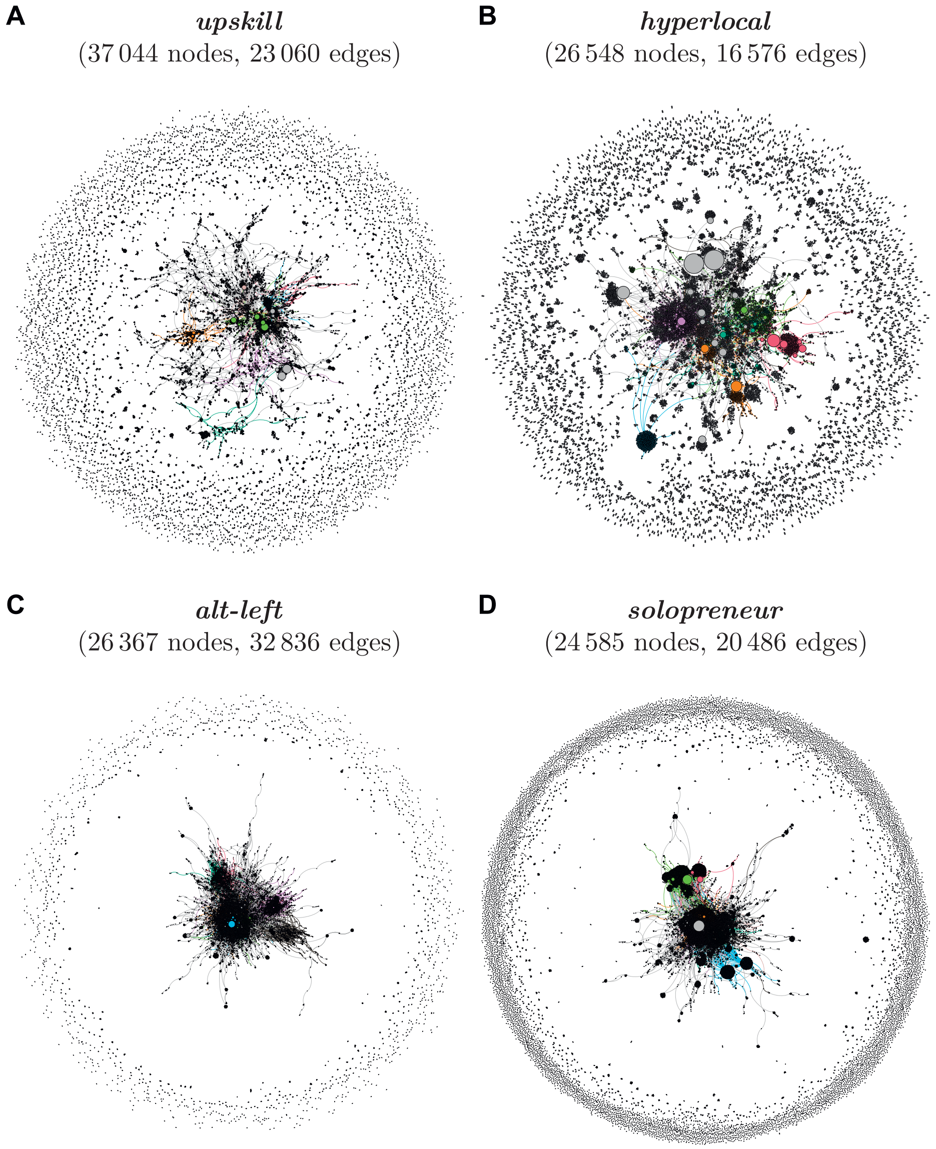

A
upskill
(37 044 nodes, 23 060 edges)
B
hyperlocal
(26 548 nodes, 16 576 edges)
C
alt-left
(26 367 nodes, 32 836 edges)
D
solopreneur
(24 585 nodes, 20 486 edges)

FIGURE 3 | Social network graphs for the last subset of the selected neologisms.

around Remembrance Day, which takes place in November. The term *poppygate* represents a last category of neologisms in the sample, which show strong fluctuations in usage intensity, but for which these patterns follow a regular temporal pattern.

To quantify the degree to which neologisms are used with consistent frequency over time, I calculate and compare the coefficients of variation for each neologism in the sample. This metric captures the overall volatility in usage frequency of words over their lifespan relative to their average frequency of occurrence in the corpus. **Tables 5–7** presents the coefficients of variation for the selected neologisms, as well as for the top and bottom six neologisms that show the highest and lowest degrees of variation in the sample.

The results in **Tables 5–7** show that the sample covers a broad spectrum of volatility in usage frequency. Among the neologisms that were used the most consistently, i.e., exhibit the lowest degrees of variation, we find words whose frequency-based measures suggested high degrees of conventionality. For example, *twitterverse* is listed among the most frequent neologisms in **Table 1** and is also one of the oldest neologisms, with its first attestation in the corpus dating back to December 19, 2006.

TABLE 5 | Coefficients of variation (VAR) for the selected neologisms.

Lexeme	VAR
hyperlocal	0.98
upskill	1.14
solopreneur	1.20
alt-right	1.81
poppygate	4.75
alt-left	5.31

TABLE 6 | Coefficients of variation (VAR) for the six neologisms with the lowest scores in the sample[6].

Lexeme	VAR
followership	0.71
lituation	0.72
twitterverse	0.72
detweet	0.74
remoaners	0.76
twittersphere	0.77

TABLE 7 | Coefficients of variation (VAR) for the six neologisms with the highest scores in the sample.

Lexeme	VAR
upskirting	9.39
youthquake	6.32
alt-left	5.31
birther	5.00
poppygate	4.75
cherpumple	4.69

TABLE 8 | Degree centrality scores (CENT) for the selected neologisms; the scores are based on the most recent time slice for each neologism in the corpus.

Lexeme	CENT
upskill	0.0021
hyperlocal	0.0085
alt-right	0.0144
alt-left	0.0238
solopreneur	0.0523
poppygate	0.0566

TABLE 9 | Degree centrality scores (CENT) for the six lexemes with the lowest scores in the sample; the scores are based on the most recent time slice for each neologism in the corpus.

Lexeme	CENT
baecation	0.0005
fleek	0.0009
ghosting	0.0013
man bun	0.0016
big dick energy	0.0018
twittersphere	0.0020

TABLE 10 | Degree centrality scores (CENT) for the six lexemes with the highest scores in the sample; the scores are based on the most recent time slice for each neologisms in the corpus.

Lexeme	CENT
rapugee	0.2580
levidrome	0.2373
kushnergate	0.2309
dronography	0.1530
dotard	0.0979
ecocide	0.0922

By contrast, the group of lexemes that show the highest degree of volatility in usage frequency is comprised of neologisms with lower degrees of conventionality, which are generally less frequent and were coined more recently. Notably, topical spikes play a crucial role in the diffusion processes of all examples in this category: the diffusion of *alt-left* and *birther*[7] was promoted by extralinguistic political events, *upskirting*[8] and *youthquake*[9] were advanced through increased metalinguistic salience after they were added to the OED and awarded Word of the Year 2017 by Oxford University Press. Both *poppygate* and *cherpumple*[10] exhibit recurrent topicality, and are typically only used in the contexts of their seasonal relevance in autumn and winter.

The selected neologisms cover the spectrum of volatility in usage frequency found in the full sample of neologisms, and the coefficients of variation represent quantitative measures which reflect the differences in volatility between the selected neologisms visualised in **Figure 2** and discussed above. The frequency-based analysis of the three neologisms discussed above demonstrates that usage frequency counts, particularly when combined with an analysis of their underlying temporal dynamics, can help to approximate the spread and success of neologisms to a certain degree. However, the results also point to substantial limitations of frequency-based approaches to studying diffusion.

The present data demonstrate considerable variation in the degrees of diffusion of neologisms with similar frequency of occurrence in the corpus. Total frequency counts alone would predict high degrees of diffusion for neologisms such as *alt-left*, for example. However, its usage history reveals that its use largely goes back to a short period of high usage intensity linked to a specific topical event. The term's background suggests that it might not have spread far beyond one particular community of speakers. Such potential distortions of frequency-based measures could partly be resolved by in-depth analyses of temporal usage profiles combined with insights from corpus data and extralinguistic events. However, these in-depth analyses of diffusion are not possible through a systematic frequency-based analysis alone, and they cannot be extended to the large-scale study of larger samples of neologisms. Hence it remains unknown to what degree frequency-based metrics adequately capture social pathways of diffusion. In the following section, I will complement the frequency-based approach by social network analyses to get a more differentiated view of the sociolinguistic aspects of diffusion.

[7] Proponent of the 'birther movement', a conspiracy theory which claims that President Obama's birth certificate was forged and that he was not born in the United States.

[8] 'The habit or practice of taking upskirt photographs or videos' (OED).

[9] 'A significant cultural, political, or social change arising from the actions or influence of young people' (https://languages.oup.com/word-of-the-year/2017/).

[10] Cherpumple is short for cherry, pumpkin and apple pie. The apple pie is baked in spice cake, the pumpkin in yellow and the cherry in white (https://en.wikipedia.org/wiki/Cherpumple); typically consumed during the holiday season in the US.

5.2 Social Networks of Diffusion

As described in **Section 4**, the social network analysis is based on the interactions between all speakers who have used the neologisms in the sample. Speakers are represented as nodes in the network graph, and interactions between users in the form of replies or mentions are represented as edges. The network structure of the resulting graphs allows analysing the degree to which the target neologisms have diffused in these networks. To monitor diffusion over time, I split the observed lifespan of each neologism into four equally-sized time slices. These time windows are marked by dashed vertical lines in **Figure 2**. I then generated network graphs for each time window for each neologism in the sample to analyse the individual pathways of diffusion over time and to compare degrees of diffusion between all neologisms in the sample.

5.2.1 Degrees of Diffusion

As discussed in **Section 4**, I mainly rely on degree centralization as a quantitative measure of diffusion. I consider increasing diffusion to be reflected by decreasing degree centralization of the graph, thus lower values of centrality indicate higher degrees of diffusion across social networks.

For example, the social graph users of a new word shows high centralization in early stages when its use is largely confined to one or few centralized clusters of speakers. When increasing diffusion extends the use of the term to new speakers and communities, the distribution of interactions in the speech community shows greater dispersion, which should be reflected by lower centrality scores for the social network of speakers.

Tables 8–10 report the degree centrality scores for the selected neologisms and for six lexemes with the highest and lowest scores in the sample

The neologisms with the lowest scores for degree centrality are also among the most frequent lexemes in the sample. Overall, frequency and centrality generally tend to produce similar results when used to assess degrees of diffusion. This shows usage frequency and social diffusion correlate, as one might expect. Notable deviations exist, however, and will be further discussed in **Section 5.3**.

Correspondingly, the neologisms with the highest centrality scores rank among the least frequent candidates in the sample. Notable trends among lexemes with high centrality scores are that they tend to be more recent (e.g., *dronography*[11]) and/or to exhibit high degrees of volatility (e.g., *ecocide*[12]). Moreover, this group includes political terms such as *Kushnergate*[13] and *rapugee* which are controversially discussed on the left and right ends of the political spectrum. For example, *rapugee* is a derogatory term which was coined after sexual assaults by refugees during New Year's Eve 2015/16 in Cologne, Germany. Previous work has shown that this term was consciously coined and propagated by a closely connected community of far-right activists to disparage refugees, and that its use on Twitter and on the Web has remained largely limited to these communities (Würschinger et al., 2016). This low degree of diffusion is reflected by the low centrality score for *rapugee*.

The following sections use network visualisations to provide a detailed, partly qualitative analysis of the diffusion for the selected cases to illustrate the social dynamics captured by the quantitative measure of centralization as an indicator of diffusion. The examples represent prototypical pathways based on centralization scores. The in-depth analysis of the social dynamics at play is guided by the detection of communities using modularity clustering (**Section 4**). The algorithm identifies the eight largest communities in each graph, visualised by colour. Moreover, I rely on the PageRank algorithm (**Section 4**) to assess the importance of users in the network, visualised by node colour. I use manual inspection of user accounts to validate and further investigate the role of these communities and influential users in the selected diffusion processes.

The centrality scores for the selected neologisms cover a broad spectrum of degrees of diffusion, as can be seen in **Table 8**. **Figure 3** presents the full network graphs for four of the selected cases to illustrate differences in the social networks of speakers which are captured by centrality scores.[14] The network graphs in **Figure 3** are sorted according to their degrees of social diffusion–as measured by centrality scores–from (a) to (d). Note that the number of nodes in each graph is very similar, differences between the visualized structure of network graphs are thus due to differences in the underlying social structure of communities rather than a mere function of differences in network size.

The neologism *upskill* exhibits the highest degree of diffusion, which is reflected by the highest degree of dispersion of nodes across the graph in **Figure 3A**. At the center of the graph, we find a relatively large cluster of speakers who are only loosely connected. Many of these speakers are connected via their affiliations to the world of business, where the term *upskill* is most commonly used. However, on the whole, the use of *upskill* is not limited to a coherent, closely-connected community. The majority of nodes appear towards the fringes and have no connections to the rest of the graph. Speakers use the term independently from each other, without being unified in their motivations to use the term by a common affiliation with a certain community of practice. The social network of *upskill* thus shows an advanced degree of diffusion.

The graph for *hyperlocal* in **Figure 3B** also shows a high degree of social diffusion, but its use depends more strongly on a central community of users. This core sub-network of speakers forms several smaller clusters which can be linked to certain domains of interest such as journalism, business, and startups, in which the term is most popular. Notably, we observe a stronger role of

[11] 'Dronography is the science, art and practice of creating durable images or video by recording light or other electromagnetic radiation by means of a drone flying around or above a certain scene (Urban Dictionary)'.

[12] 'the destruction of large areas of the natural environment as a consequence of human activity (Merriam Webster Online Dictionary)'.

[13] Referring to a political scandal involving Trump's senior adviser Jared Kushner allegedly meeting Russian officials.

[14] The network graphs for *alt-right* and *poppygate* were omitted as their difference in network size does not allow for comparative analyses (*alt-right*: 2,74,686 nodes, *poppygate*: 2473 nodes).

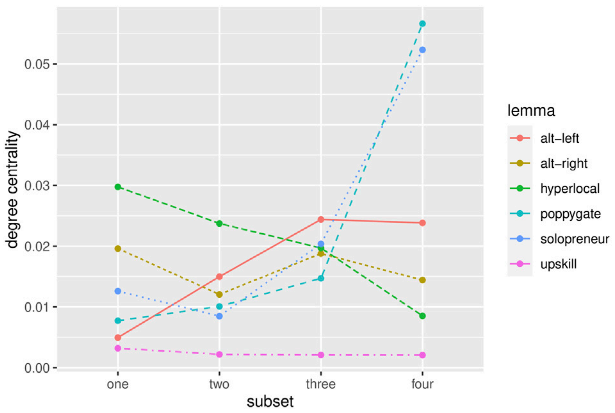


FIGURE 4 | Pathways of diffusion for the selected neologisms. The graph shows DEGREE CENTRALITY scores over time, each SUBSET representing one network graph which was generated for each of the four equally-sized time slices for each neologism in the sample.

individual user accounts such as influencers and marketing agencies, which is illustrated by bigger node sizes (representing high PageRank scores). Yet, as in the graph for *upskill*, the majority of occurrences of *hyperlocal* can be traced back to a large number of speakers from a diverse set of sub-communities, which can be interpreted as a sign of advanced diffusion.

The social graph for *alt-left* shows very limited diffusion of the term. Almost all of its use can be traced back to one closely-connected community of users. This core community of users demonstrates typical characteristics of an echo chamber in that it is dense and features strong ties within the community, but has few weak ties connecting it to the rest of the social graph. This observation is in line with the socio-political background of the term, which was coined and propagated by far-right activists in an attempt to unify political efforts ('*Unite* the Right Rally') and to distance themselves from and protest against the political left. Inspection of the network reveals that the most influential node in the network is Donald Trump. His use of the term was followed by a sharp increase in usage intensity in the course of the Charlottesville Rally in August 2017. The high degree of social compartmentalization in the use of *alt-left* is also reflected in the ratio between the number of nodes and edges in its graph, which confirms that its community of speakers is much more closely connected than that of the remaining neologisms[15]. Notably, the same applies to the community of *alt-right*, which occupies the opposite pole of the political spectrum. The results for these two terms are in line with previous work reporting effects of political polarization in online social networks for these political communities (Sunstein. 2018). Overall, *alt-left* thus shows a low degree of diffusion. It has received significant popularity in certain parts of the speech community, but its use remains strongly limited to these communities.

Lastly, the social network of speakers using the term *solopreneur* also shows limited diffusion. A significant proportion of its use comes from a diverse set of individual speakers and micro-communities, which are placed at the fringes of the graph. However, similar to the social graph for *alt-left*, a relatively well-connected, large core of speakers is responsible for the majority of its use in the corpus. Moreover, unlike the example of *alt-left*, this central community of users is in turn dominated by the high centrality of a small number of individual accounts. Inspecting the network of users reveals that these 'influencers' are all either proud, self-proclaimed solopreneurs, or coaches and agencies that are using the term to promote their services to aspiring entrepreneurs. Overall, *solopreneur* has achieved significant popularity within certain communities, but its use in these communities is unevenly distributed and depends strongly on a small number of individual users. The term does not show signs of advanced diffusion since its use is largely limited to certain individual speakers and communities of practice.

In summary, the social networks of speakers reveal significant differences in the degrees of social diffusion for the neologisms in the present dataset, as observed in the period leading up to the cutoff point at the end of 2018.

While the centrality measures generally concur with the frequency-based analysis of the neologisms discussed in **Section 5.1**, the network metrics and visualisation add information by providing a more detailed picture of degrees of social diffusion and highlight cases for which the social dynamics of diffusion diverge from what could be observed by relying on usage frequency alone.

5.2.2 Pathways of Diffusion

To investigate the pathways of social diffusion, **Figure 4** presents the degree centrality scores for the selected neologisms over time. The scores for Subset 4 represent the final degrees of diffusion as presented in **Table 8**. The corresponding network graphs for this stage were presented in **Figure 3**. The centrality scores for the preceding subsets now add information about the diffusion history of these neologisms. The diverging trajectories of centralization over time indicate significant changes over time as well as differences in the pathways of diffusion between neologisms.

[15] The numbers of edges per node for all selected cases in descending order: *alt-right*: 1.49, *alt-left*: 1.24, *solopreneur*: 0.83, *hyperlocal*: 0.62, *upskill*: 0.62, *poppygate*: 0.53.

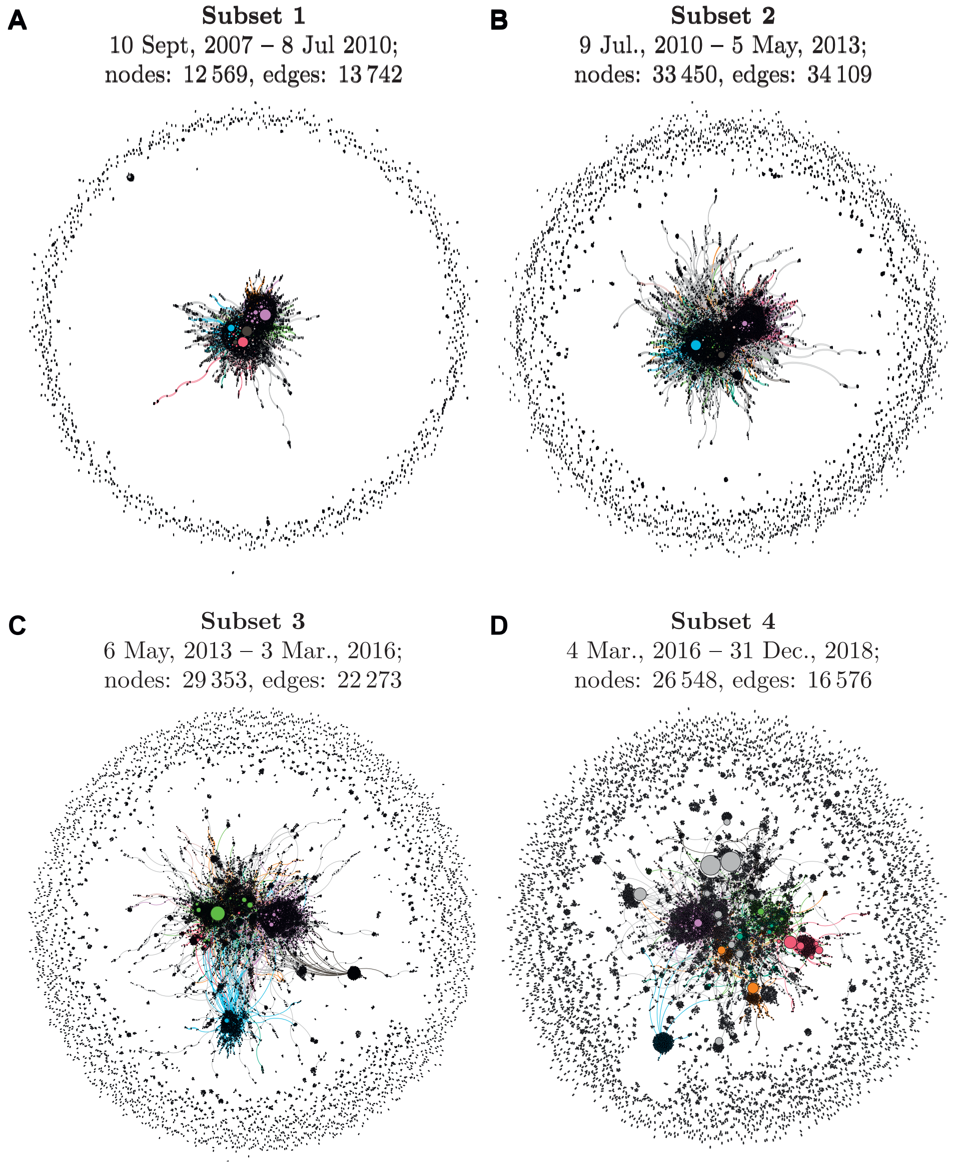


FIGURE 5 | Social network of diffusion for *hyperlocal* over time.

Figure 5 presents the full network graphs for all stages of diffusion for the term *hyperlocal* to illustrate the social dynamics underlying the quantitative measures.

Both the quantitative measure in **Figure 4** and the network visualizations in **Figure 5** indicate that *hyperlocal* shows increasing, successful diffusion over time. Its use is relatively centralized in its earlier stages, which can be seen from the fact that most speakers who have used the term are closely connected in the social graph in the first quarter of its observed lifespan. Inspecting the most influential speakers and sub-communities in the network (based on PageRank and Modularity scores) reveals that *hyperlocal* is mainly used by a relatively small community of individual journalists in the first subset, who are early adopters in trying to target news to local audiences and use the term very frequently to label this new approach.

In Subset 2, the community of journalists grows and starts to include also bigger news outlets such as *The Guardian*. Additionally, a new community of practice adopts the term: several marketing agencies start promoting their services using the term *hyperlocal*. At this point, the usage intensity of the term peaks, as was demonstrated in **Figure 3B**. However, the social network data indicate that at this point its use is still mainly the product of high popularity and usage intensity within a small number of dense sub-communities rather than a sign of advanced diffusion across bigger parts of the speech community.

The network graphs show that the social diffusion of *hyperlocal* is only significantly advanced in the last two stages.

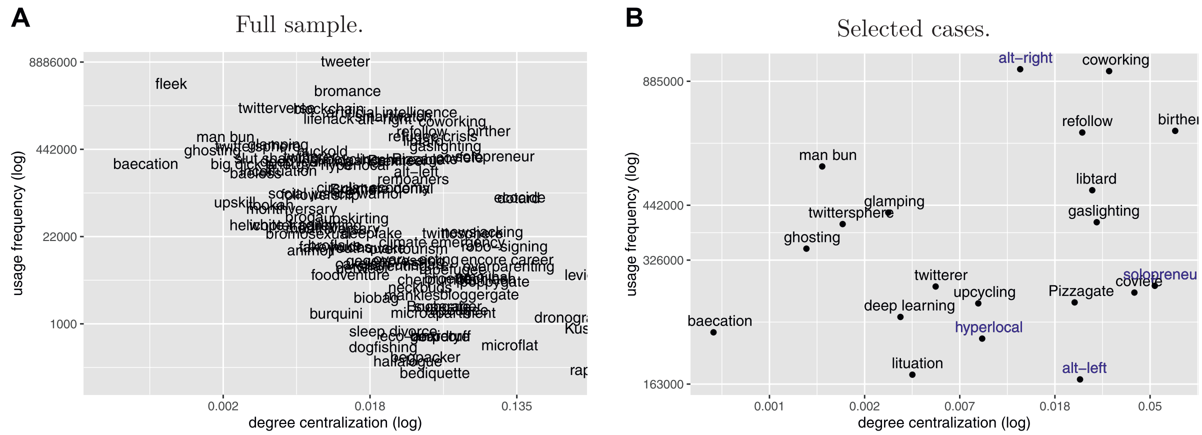


FIGURE 6 | Relationship between total USAGE FREQUENCY and degree centrality (CENTRALIZATION) for the full sample of neologisms (n = 99) and the selected cases.

TABLE 11 | Correlations of 'degree centralization' (CENTRALITY) with the variables total usage frequency (FREQUENCY), coefficient of variation (VOLATILITY), and observed lifespan in the corpus (AGE) for the full sample of neologisms (n = 99) using Spearman's correlation coefficient (Spearman 1961)[17].

	ρ	p
Frequency	−0.44	<0.001
Age	−0.29	0.004
Volatility	0.28	<0.001

While we see only few weak ties during the earlier stages of its use, the term now increasingly diffuses beyond its early adopters. Inspecting the network reveals that the use of the term becomes increasingly popular in the world of business and startups as well as the general public on Twitter. The network metrics indicate that individual agents and sub-communities now play a far smaller role in its overall use. While *hyperlocal* shows less usage intensity during these later stages, the network metrics indicate a high degree of diffusion for the second half of its observed lifespan. The timing of its addition to the OED in 2015 supports these observations. The term *hyperlocal* has successfully spread beyond its subcommunities of early adopters, and it seems to be used by a diverse community of speakers from different backgrounds, which renders it a case of advanced diffusion. This process of increasing diffusion for *hyperlocal* is also reflected in its decreasing measures for graph centrality in **Figure 4**.

The remaining cases in **Figure 4** show different pathways of diffusion, both in terms of their overall degree of diffusion and diachronic trajectory. Due to space limitations, I can only provide an overview of their development over time.

Besides *hyperlocal*, the second neologism which exhibits advanced diffusion is *upskill*. In this case, however, we observe little change over time, its degree centrality has been very low since its early attestations in the corpus. This indicates a gradual spread across speakers which is not significantly affected by a small group of influential speakers. The term *upskill* has been used by a wide variety of speakers throughout its observed lifespan and shows the highest degree of diffusion among the selected cases.

By contrast, *solopreneur* and *poppygate* show a negative trend in terms of diffusion. The term *solopreneur* features low degrees of diffusion in its earlier stages, but its use becomes more centralized over time. This is in contrast with its usage intensity over time (**Figure 2**): while its earlier period of moderate use goes back to a decentralized cluster of users, its increase in usage frequency coincides with a narrowing of its user base. As the network analysis in **Figure 3D** demonstrates, it becomes increasingly limited to a relatively small community which shares interest in a small professional niche.

The case of *poppygate* exhibits a similar trend towards increasing centralization. Its temporal dynamics show a pattern or recurrent topical usage (**Figure 2**). The social networks of *poppygate* suggest that while the term was used by a broader audience in its earlier stages, its use in the more recent past goes back to certain communities of speakers for which a specific topical event emerges as a salient occasion to use the term. For example, its most recent spike in usage intensity in November 2016 was caused by a controversy about whether Fifa was right to take disciplinary action against the national teams of England and Scotland after their players wore poppy armbands during a football match between the two nations on 11 November. Protests by the football community caused a spike in usage intensity for *poppygate*, but did not trigger its diffusion beyond this community[17].

Lastly, *alt-right* and *alt-left* show limited degrees of diffusion over their lifespan. While the centrality of *alt-right* remains fairly stable over time, *alt-left* shows increasing centralization. Both terms are strongly tied to the political discourse surrounding the Unite the Right Rally in the United States and consequently exhibit a sharp increase in usage intensity in the course of the event in August 2017 (**Figure 2**). This increase in use is, however, reflected by increased centrality

[17] All variables entering the correlation analysis were log-transformed and centred. I report Spearman's correlation coefficients to avoid assumptions about the linearity of the variables involved. I additionally calculated Pearson's correlation coefficients, for which the correlation coefficients are slightly higher: FREQUENCY: $\rho = -0.45$, $p < 0.001$; AGE: $\rho = -0.38$, $p < 0.001$; VOLATILITY: $\rho = 0.23$, $p < 0.001$.

scores for both lexemes in **Figure 4**. This period of highly intense use is thus characterised by relatively smaller rather than larger degrees of diffusion for both lexemes. While the use of *alt-right* reverts to more decentralized use afterwards, the use of *alt-left* remains at this high level of centrality. This seems to confirm the echo chamber effect for *alt-left* discussed in **Section 5.2.1**: the term has become conventional and popular among a community of like-minded individuals, but its use remains limited to this community. Given the extreme, far-right attitudes and political orientations prevalent in this group, the majority of Twitter users do not want to be associated with this community of users. Since the term *alt-left* has become highly indexical of support and membership of this political camp, very few speakers are willing to adopt and use the term.

In summary, studying the temporal dynamics of social networks highlights changes in the use of neologisms over time and reveals differenct pathways of diffusion in the sample.

5.3 Combining Frequency and Network Information

Having applied the frequency-based and the social network approach to assess the diffusion of the present sample of neologism, this section will combine the results obtained from both approaches and show how they complement each other[16].

5.3.1 Correlations

A first evaluation of the social network approach to diffusion relies on the correlations of degree centrality with the total usage frequency of neologisms, with their volatility, and with their age as observed in the corpus. **Table 11** reports the correlation coefficients for these variables.

Firstly, centrality shows a significant negative correlation with FREQUENCY. This confirms earlier observations in **Section 5.2** which indicated an inverse trend between total usage frequency and centrality. More frequent neologisms show on average higher degrees of diffusion, i.e. increase in frequency correlates with wider spread across the speech community. The fact these two central measures for diffusion correlate can be seen as a cross-validation of both approaches. While external data sources would be needed for a more rigorous evaluation, this overall convergence in results suggests that both metrics capture important aspects of diffusion.

Secondly, the AGE of neologisms in the sample shows a significant negative correlation with centrality. As expected, the use of more recent neologisms tends to still go back to more centralized communities, while neologisms with a longer history of use tend to show more advanced diffusion. Unlike frequency counts, which are directly influenced by the temporal usage history of neologisms, the centrality measure is blind to this information. The fact that these age effects are captured by degree centrality supports the usefulness of the social network approach.

Lastly, VOLATILITY shows a significant positive correlation with centrality. Again, this result is in line with expectations. Neologisms such as *poppygate*, whose use exhibits substantial temporal variation tend to show lower degrees of diffusion than neologisms such as *hyperlocal*, whose use is more consistent and less dependent on the topical salience of extralinguistic events.

5.3.2 Deviations Between Centrality and Frequency

For a closer analysis of the interactions between these variables beyond correlation coefficients, **Figure 6** presents all neologisms according to their usage frequency and centrality scores. While **Figure 6A** covers the full sample, **Figure 6B** is based on the same data, but zooms in on the frequency range which covers four of the selected cases to provide a clearer view of this section of the sample.

The general trend in the plot confirms the inverse relation captured by the negative correlation coefficient between centrality and frequency. Neologisms with high frequency such as *fleek* have low centrality scores and would thus be assigned a high degree of diffusion by both approaches. The inverse applies to candidates from the lower end of the frequency spectrum such as *microflat*.

However, **Figure 6A** also shows substantial variation between frequency and centrality scores. Notably, the observed deviations are almost exclusively found towards the right of the diagonal trend, i.e., for cases where centrality assumes lower degrees of diffusion than frequency. For example, while *fleek* and *bromance* are assigned similar scores in terms of their usage frequency, their centrality scores suggest a much lower degree of diffusion for the latter neologism. Similar to cases like *solopreneur* and *alt-left*, which were discussed in detail in **Section 5.2.1**, centrality thus provides additional information for cases in which the social network structure indicates that the observed usage intensity overestimates the degree of diffusion of a target neologism. This can arise if its observed uses go back to a disproportionately smaller number of speakers and subcommunities.

Analysing these deviations highlights two main groups among the selected neologisms, for which total usage frequency and social network structure seem to diverge in systematic ways[18]. A first group contains neologisms marked by high degrees of volatility in their frequency of use. As shown above, centrality is significantly correlated with volatility. In addition to *poppygate* and *solopreneur*, which were already discussed above, *refollow*, *gaslighting*, *solopreneur*, and *coworking* also show little consistency in their usage. For all of these terms, social diffusion is out of sync with the increase in usage intensity in **Figure 6A**. It thus seems that the social network approach adds an extra layer of information which comes to the fore especially where frequency-based measures overestimate degrees of diffusion due to the strong impact of short periods of highly intensive use of neologisms in certain parts of the speech community.

[16] It should be noted that a strict evaluation of both approaches is in principle impossible without external data about the degrees of diffusion for the neologisms under investigation. While such a gold standard for evaluation is inconceivable in the present context, it would be desirable to use additional data sources such as questionnaires, dictionaries or web corpus data for a more rigorous validation of the present approach. This will have to be left for future work.

[18] The present dataset does not allow to assess whether the deviations of the two groups that emerge in this analysis are generalisable.

A second, converse group with diverging scores contains neologisms whose use is tied to political communities. The neologisms *alt-right*, *alt-left*, *birther*, *covfefe*, *Pizzagate*, and *Kushnergate* are politically controversial and differ strongly in popularity between political camps. It should be noted that these terms also exhibit considerable volatility in their use. **Figure 6A** shows comparatively lower centrality than frequency scores for these lexemes. Similarly to the cases of high volatility, centrality thus suggests that usage frequency overestimates degrees of diffusion for these cases. While neologisms such as *alt-right* show high frequency counts, the social network analysis reveals that these terms have not spread successfully across communities, and that their use remains limited to certain subcommunities.

5.3.3 Predicting the Success of Lexical Innovations

The results from the network approach show that community structure can be used to assess degrees of diffusion. The social structure of communities during the early stages of diffusion is commonly assumed to be an important factor for the successful spread of linguistic innovations. While a detailed analysis is beyond the scope of the present paper, the present approach yields initial results of the predictive power of social network information.

The dataset shows a significant correlation between the network structure in the first period of diffusion and the overall success of neologisms. Correlating CENTRALITY scores for all neologisms in Subset 1 with their total usage FREQUENCY observed across their full observed lifespan in the corpus yields Spearman correlation coefficient of −0.43 ($p < 0.001$). This means that neologisms are overall more likely to spread successfully if their use is not limited to a centralized network of speakers in their early stages. Among the selected cases presented above, *upskill* fits this pattern: it shows a consistent, successful trajectory of diffusion and its use has been the product of a decentralized bunch of users since its early attestations. Of course, the diverging pathways of diffusion for other words such as *hyperlocal* and *solopreneur* presented in **Figure 4** represent exceptions to this general trend. While this trend fits theoretical expectations and the empirical observations in the present dataset, these results remain preliminary. Since centrality correlates with frequency scores, future work based on larger samples, external data for evaluation, and more robust statistical tests is needed to test whether the predictive power of social network features can be confirmed.

6 DISCUSSION

In this paper, I have studied the spread of neologisms on Twitter to provide a multi-layered picture of the diffusion of lexical innovations in terms of 1) overall usage frequency, 2) changes in usage frequency over time (volatility), and 3) pathways of social diffusion across members and networks in a larger speech community. The process of diffusion entails social processes which lead to the spread of innovations in social networks (Rogers 1962). Theoretical models characterise the spread of linguistic innovations to new speakers and communities as the key feature of the process of diffusion (Weinreich et al., 1968; Schmid 2020). Despite a broad consensus over the fact that diffusion entails spread in networks of speakers, most previous empirical investigations of lexical innovation have not been based on social network information, but have relied on frequency measures as an indicator for the diffusion of neologisms (Stefanowitsch and Flach 2017). The present study used a large Twitter dataset to investigate the sociolinguistic dynamics of diffusion of neologisms in online social networks. Aside from an in-depth analysis of the spread of neologisms in the present sample, the aim of this paper was to assess the usefulness of using usage frequency and social network data as indicators of diffusion.

6.1 Temporal Dynamics of Diffusion

The frequency-based approach revealed that frequency measures can be used to assess degrees of diffusion of lexical innovations with varying success. Total frequency counts (**Tables 1–4**) proved successful for a coarse-grained distinction between cases of high (e.g., *tweeter*, *smartwatch*), medium (e.g., *monthiversary*, *helicopter parenting*), and low degrees of diffusion (e.g., *begpacker*, *bediquette*). However, differences in the temporal dynamics of use have proved to be necessary for a more accurate assessment of the degrees and pathways of diffusion of neologisms.

Considering the nature of the process and products of *lexical* innovation, this temporal sensitivity is not surprising. Models of linguistic diffusion such as the S-curve model assume competition processes in which several formal variants compete to become the conventional linguistic means to express a certain meaning/function in the speech community. In cases of grammatical innovation, which is at the core of most models and most previous empirical investigations of diffusion, the communicative need for expressing the target concept/function remains stable over time. While grammatical means are, of course, also subject to language change (e.g., *going to*, *will* future), the salience of the target semasiological space (e.g., 'expressing future intention'), remains stable over time for all speakers in the speech community. Both the direct competition between linguistic variants and the social and temporal invariance of the conceptual space over time are tacit assumptions of S-curve models of diffusion (Blythe and Croft 2012).

Earlier work by Nini et al. (2017) suggests that the diffusion of lexical innovations also follows S-curve trajectories, and the authors use the term 'semantic carrying capacity' to refer to the semantic potential of neologisms during diffusion. It seems plausible that the semantic carrying capacity of new words exhibits significant volatility over time and across communities of speakers. While the present study cannot measure or control for changes in semantic potential over time, it tries to account for the temporal sensitivity of neologisms by going beyond cumulated frequency counts and studying their temporal usage profiles.

The present study focused on three main aspects of the temporal dynamics of diffusion: trends in usage intensity, age and volatility. Firstly, trends in usage frequency add information about changes in the degrees of diffusion of neologisms over time. Going beyond total frequency counts, visualising the cumulative increases in usage frequency over time in **Figure 1** revealed significant differences in the pathways of diffusion of neologisms with similar total frequency counts. The neologism *hyperlocal* showed the most linear trajectory indicating fairly consistent use, the convex curve of *upskill* indicated a positive trend in its use, and the concave trajectories of *solopreneur* and *alt-left* suggested negative trends in the recent past.

Cumulated frequency counts, which are, in their pure form as total counts, agnostic to temporal trends, have successfully been used as an approximation of the 'potential exposure' (Stefanowitsch and Flach 2017) of speakers to linguistic constructions in previous usage-based corpus-linguistic studies. The present results emphasize, however, that temporal trends and changes in usage frequency cannot be neglected when assessing the social diffusion of neologisms, since innovation in the lexicon is subject to high degrees of temporal variation. Notably, trends in usage frequency in the present sample can almost always be traced back to changes in the neologisms' semantic carrying capacity and are not merely the product of onomasiological competition between formal variants[19]. Typical examples of the influence of topical salience on the use of neologisms are re-current topical neologisms like *poppygate* discussed in **Section 5.1.1**.

Secondly, it was shown that the age of neologisms provides important information about their diffusion processes. Neologisms such as *hyperlocal* and *alt-left*, which are comparable in total use frequency, but differ strongly with regard to their observed lifespan in the corpus, show different pathways and degrees of diffusion. Older neologisms whose use is distributed more evenly across longer periods of consistent usage (*hyperlocal*) typically show higher degrees of social diffusion than younger neologisms whose use almost exclusively goes back to a short period of highly intensive use (*alt-left*). The positive relationship between the age of neologisms and their degrees of diffusion was supported by the significant correlation with centrality in the network analysis. While a longitudinal, predictive approach to the fate of lexical innovations is beyond the scope of the present paper, it seems possible that neologisms follow Lindy's Law: the longer new words have been in use in the speech community, the less likely they are to become obsolete in the (near) future (Eliazar 2017). The fate of new words ultimately depends on the conceptual salience of the objects and practices they denote, however: whether *smartwatch* and *blockchain* outlive previous neologisms such as *Walkman* and *Discman* ultimately depends on the future success of these products in our society.

Lastly, the results showed that volatility in use is an important factor in the diffusion of neologisms. While some candidates show fairly consistent usage frequency over time (e.g., *hyperlocal*, *upskill*), most exhibit considerable fluctuations. For some words in the sample, recurrent spikes in usage intensity are an inherent part of their usage profile. The neologism *youthquake* is characterised by spikes in usage intensity when relevant to current public affairs, but shows low frequency of use in the intermediate intervals. Due to the nature of this behaviour, this pattern has been termed 'topical' by Fischer (1998). Cases such as *poppygate*, for which these topical spikes occur in fairly regular, periodic intervals, have been classified as 'recurrent semi-conventionalization' by Kerremans (2015). For both groups of neologisms total frequency counts cannot provide an accurate estimation of degrees of diffusion since they lack information about these patterns of volatility which are central to these cases of lexical innovation. The network approach to diffusion in **Section 5.2** revealed a negative correlation between volatility and degrees of diffusion. It seems that neologisms that are used less consistently over time are less likely to reach advanced degrees of diffusion. Moreover, comparing frequency counts and degree centrality indicated that frequency tends to overestimate the degree of diffusion of topical neologisms. This is in accordance with the observation that isolated spikes in usage intensity tend to go back to disproportionally smaller parts of the speech community.

6.2 Social Dynamics of Diffusion

To get a more differentiated view of the social dynamics of diffusion, I conducted a social network analysis of the present dataset. Successful diffusion was defined in **Section 2** as spread to new speakers and new communities. Unlike measures such as frequency and volatility which are solely based on the occurrence of neologisms in the corpus, the network approach is based on the social structure of the networks of speakers who have used the target neologisms and thus provides a more direct operationalisation of social pathways of diffusion.

The present results show considerable overlap between frequency and network measures of diffusion. Network centrality significantly correlates with usage frequency, and visualising the relationship between both metrics (**Figure 6A**) confirms this trend. Both metrics assign high scores for diffusion to established neologisms such as *man bun*, and low scores to less established candidates such as *microflat*. Moreover, centrality shows significant correlations with age and volatility, thus confirming the intuition and general finding that higher usage intensity correlates with wider social diffusion.

The more detailed evaluation of both approaches in **Section 5.3.2** also revealed that usage frequency is an imperfect predictor of social diffusion. Centrality generally tends to assign lower degrees of diffusion than frequency for some of the cases in the sample. The main groups affected consist of neologisms whose use goes largely back to specific communities of practice (e.g., *solopreneur*), political communities (e.g., *alt-left*), and/or highly volatile neologisms (e.g., *poppygate*). A closer analysis of these cases in **Section 5.2** showed that in these cases the observed number of uses of these neologisms stems from a comparatively smaller number of speakers and communities. It thus seems that the social network information contained in the measure of centrality manages to account for cases in which total usage frequency overestimates degrees of diffusion.

These discrepancies in results reflect two perspective on the process diffusion. Successful diffusion of neologisms was defined as spread to new speakers and new communities. Using the frequency of occurrence of a neologism in a corpus to approximate to what degree it is familiar to bigger parts of the speech community thus has to rely on several assumptions which are only accurate to a certain extent.

Firstly, the number of uses observed might diverge from the number of speakers who are familiar with the term. Frequency can overestimate the latter, for example, if the observed use is the product of high usage intensity by a smaller number of speakers (e.g., *solopreneur*) rather than moderate use by a higher number of speakers (e.g., *hyperlocal*).

[19] As an exception, the sample contains two sets of formal variants: *monthversary* & *monthiversary* and *rapefugee*, *rapeugee* & *rapugee*.

Secondly, usage frequency only captures active uses of the term and is blind to the number of speakers who are familiar with the term, but have not used it in the corpus. By contrast, social network metrics also include speakers who have only been passively exposed to the term, and thus covers a broader, and arguably more relevant definition of 'familiarity'. Network metrics are free from the assumption that the observed output of speakers in the corpus is representative of the input to speakers in the speech community (Stefanowitsch and Flach 2017).

Lastly, the number of uses observed might not be indicative of whether a neologism has spread beyond certain sub-communities and has reached a broader spectrum of the speech community. Many of the neologisms for which centrality indicates significantly lower degrees of diffusion than frequency are socio-politically loaded and known to be used by fragmented and polarized communities, mainly from the far-right end of the political spectrum (Sunstein. 2018). **Figure 6B** features terms such as *alt-right*, *alt-left*, *birther*, *covfefe*, *Pizzagate*, and *Kushnergate*. Among the selected cases, *alt-left* and *hyperlocal* show a similar total number of uses. Moreover, the numbers of users involved in its use in the last temporal subset are almost identical: 26,367 vs. 26,548. Yet, their social network structure in **Figure 3** and their centrality scores indicate far lower degrees of diffusion for *alt-left*. While this political term has become popular among a closely connected community of users, its conventionality remains limited to this social niche and does not extend to bigger parts of the speech community. Its isolated use is in accordance with the socio-linguistic background of the term which was consciously coined by far-right activists as a disparaging out-group term in an attempt to 'Unite the Right'.

The potential distortions that may arise when assessing the degrees of conventionality of linguistic constructions on the basis of usage frequency alone apply in principle to all linguistic domains. However, the underlying assumptions are particularly problematic in the case of lexical innovation.

Firstly, linguistic *innovations* are by definition new and not (yet) conventional among the speech community. It is therefore to be expected that their use is unevenly distributed across communities of speakers. Since frequency counts alone do not provide information about this distribution, sociolinguistic data are needed to assess the degrees of social diffusion of linguistic innovations.

Secondly, unlike linguistic innovations in other domains such as morphology or syntax, *lexical* innovations are often consciously coined and have a very specific communicative function. Their usefulness is closely tied to the conceptual salience of the entity they denote. The semantic carrying capacity of new words is thus much more likely to exhibit social and temporal variation than the functional potential of grammatical constructions. While speakers of English from all walks of life have felt the urge to talk about the future, the urge to talk about the future of 'blockchain' has only come up very recently, is (still) limited to specific parts of the speech community, and might not persist in the future. In other words, the use of lexical innovations exhibits greater social and temporal variation than innovations in other linguistic domains. The interpretation of aggregated frequency counts, which suggest a uniform distribution of use across time and across the speech community, is thus particularly problematic for assessing the diffusion of new words.

Moreover, neologisms typically arise in specific communities of practice and often show, at least initially, high degrees of social indexicality with regard to these communities. The present dataset includes several neologisms which are associated with youth language (*fleek*, *lituation*) and political discourse (*birther*, *alt-left*), for example. A term like *alt-left*, which could in principle be used neutrally to designate the political far-left, is highly socially indexical of the far-right community it emerged from. Therefore it is less likely to be used by speakers outside this community, unless they are willing to be associated with this community. Neologisms which are socially indexical are thus more community-specific. Even when speakers outside this community are familiar with these terms, they are less likely to use them. Usage frequency counts miss such effects, since they only capture active uses of neologisms.

7 CONCLUSION

In summary, the present study has shown that frequency and network-based approaches capture different kinds of information about the use and spread of new words. As we have seen, both approaches show considerable overlap in their overall assessment of degrees of diffusion. On the one hand, measures which are based on the occurrence of neologisms in the corpus such as frequency, age, and volatility capture important aspects about the temporal usage profiles of neologisms. On the other hand, social networks provide a more differentiated view of the social dynamics of diffusion. They allow to visualise and quantify different pathways and degrees of diffusion, which enables a more detailed analysis of the spread of new words to new speakers and communities. While the approaches differ in their strengths and weaknesses, combining information from both approaches provides the most complete picture of diffusion, of course. In corpus-linguistic practice, total frequency counts are the most readily available and most widely used measure for the conventionality of linguistic constructions. The present results suggest that the additional consideration of temporal dynamics of use and social network information can contribute substantially towards a more detailed and accurate picture of diffusion.

As I have argued, the use of network information is of particular importance for the study of neologisms, due to the nature of the process of lexical innovation. However, social network analysis also has great potential for sociolinguistic research in other domains. One of its biggest advantages is that it is usage-based and captures the communicative behaviour of speakers in interaction. It thus enables very fine-grained analyses of the sociolinguistic dynamics of communities, which can be visualised and qualitatively inspected on the basis of network graphs. Additionally, network science offers powerful algorithms to quantify and model the social characteristics of communities on a macro level.

The interactional dynamics discovered by network analyses can be a valuable addition to more traditional, static sociolinguistic information such as metadata about groups of speakers. Moreover, network analyses can be used in cases where metadata about speakers are unavailable, as in the present study. Since the importance of online social networks like Twitter and Reddit is only going to grow in the future, both in terms of their role in society and in academic research, network analyses have great potential for future sociolinguistic research.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Twitter's Terms of Service. Requests to access these datasets should be directed to QW, q.wuerschinger@lmu.de.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I would like to thank Stefano Coretta, Cornelius Fritz, Hans-Jörg Schmid, Maximilian Weigert, and the two reviewers for their comments on this paper.

REFERENCES

Banes, K. (2014). *Free Tools for Writers, Bloggers and Solopreneurs*. Seattle, Washington: Amazon. Available at: https: //www.amazon.com/-/de/gp/product/B00IOW2QI0.

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks," in Third International AAAI Conference on Weblogs and Social Media, March, 2009, San Jose, CA. Available at: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., and Dodds, P. S. (2012). Twitter Reciprocal Reply Networks Exhibit Assortativity with Respect to Happiness. *J. Comput. Sci.* 3 (5), 388–397. Advanced Computing Solutions for Health Care and Medicine. Available at: http://www.sciencedirect.com/science/article/pii/S187775031200049X. doi:10.1016/j.jocs.2012.05.001

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Blythe, R. A., and Croft, W. (2012). S-curves and the Mechanisms of Propagation in Language Change. *Language* 88 (2), 269–304. doi:10.1353/lan.2012.0027

Brin, S., and Page, L. (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in Seventh International World-Wide Web Conference (WWW 1998), April, 1998, Brisbane, Australia. Available at: http://ilpubs.stanford.edu:8090/361/. doi:10.1016/s0169-7552(98)00110-x

Bruns, A. (2012). How Long Is a Tweet? Mapping Dynamic Conversation Networks Ontwitterusing Gawk and Gephi. *Inf. Commun. Soc.* 15 (9), 1323–1351. doi:10.1080/1369118X.2011.635214

Camenisch, J., Lambrinoudakis, C., Jódar, L., Cortés, J. C., and Acedo, L. (2011). Public Key Services and EUROPKI-2010-Mathematical Modelling in Engineering & Human Behaviour. *Math. Comp. Model.* 57 (7), 1577–2028. Available at: https: /www.sciencedirect.com/science/article/pii/S0895717711007898 (Accessed 02 06, 2021).

Cartier, E. (2017). "Neoveille, a Web Platform for Neologism Tracking," in Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, March, 2007, (Valencia, Spain: Association for Computational Linguistics), 95–98. Available at: https://aclweb.org/anthology/E17-3024. doi:10.18653/v1/e17-3024

Davies, M. (2013). Corpus of News on the Web (NOW) - 3+ Billion Words from 20 Countries. Available at: https://www.english-corpora.org/now/ Accessed August 6, 2021.

Del Tredici, M., and Fernández, Rl. (2018). "The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities," in Proceedings of the 27th International Conference on Computational Linguistics. Available at: https://arxiv.org/abs/1806.05838.

Dunbar, R. I. M. (1992). Neocortex Size as a Constraint on Group Size in Primates. *J. Hum. Evol.* 22 (6), 469–493. doi:10.1016/0047-2484(92)90081-j

Eisenstein, J., O'Connor, B., Smith, N. A. P., and Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLOS ONE* 9 (11), e113114–13. doi:10.1371/journal.pone.0113114

Eliazar, I. (2017). 'Lindy's Law'. *Physica A: Stat. Mech. its Appl.* 486, 797–805. doi:10.1016/j.physa.2017.05.077

Elsen, H. (2004). *Neologismen. Formen Und Funktionen Neuer Wörter in Verschiedenen Varietäten Des Deutschen*. Tübingen: Narr.

Fischer, R. (1998). *Lexical Change in Present Day English. A Corpus Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Narr.

Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks* 1 (3), 215–239. Availble at: http://www.sciencedirect.com/science/article/pii/0378873378900217 (visited on 02/06/2020. doi:10.1016/0378-8733(78)90021-7

Gérard, C., Bruneau, L., Falk, I., Bernhard, D., and Rosio, A.-L. (2017). 'Le Logoscope : Observatoire Des Innovations Lexicales En Français Contemporain'. In: *La Neología En Laslenguas Románicas: Recursos, Estrategias Y Nuevas Orientaciones*, Editor J. Palacios, G. de Sterck, D. Linder, J. del Rey, M. S. Ibanez, and N. M. García. Frankfurt a. M., Germany: Peter Lang. Availble at: https://hal.archives-ouvertes.fr/hal-01388255.

Gerlitz, C., and Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C J.* 16. 2013 Availble at: http://www.journal.mediaculture.org.au/index.php/mcjournal/article/view/620. doi:10.5204/mcj.620

Goel, R., Soni, S., Goyal, N., Paparrizo, J., Wallach, H., Diaz, F., et al. (2016). 'The Social Dynamics of Language Change in Online Networks'. In: *Social Informatics*. Editors E. Spiro and Y.-Y. Ahn. Cham: Springer International Publishing, 41–57. doi:10.1007/978-3-319-47880-7_3

Granovetter, M. S. (1973). 'The Strength of Weak Ties. *Am. J. Sociol.* 78 (6), 1360–1380. doi:10.1086/225469

Grieve, J. (2018). "Natural Selection in the Modern English Lexicon," in Proceedings of EVOLANG XII (Poland: Torun). doi:10.12775/3991-1.037

Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping Lexical Dialect Variation in British English Using Twitter. *Front. Artif. Intell.* 2, 11, 2019. Available at: https://www.frontiersin.org/article/10.3389/frai.2019.00011. doi:10.3389/frai.2019.00011

Grieve, J., Nini, A., and Guo, D. (2016). Analyzing Lexical Emergence in Modern American English Online. *English Lang. Linguistics.* 21, 99–127. doi:10.1017/S1360674316000526

Grieve, J., Nini, A., and Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *J. English Linguistics.* 46 (4), 293–319. doi:10.1017/s1360674316000113

Halu, A., Mondragón, R. J., Panzarasa, P., and Bianconi, G. (2013). 'Multiplex PageRank'. *PLOS ONE* 8 (10), e78293, 2013 . Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078293 (Accessed 02 06, 2021). doi:10.1371/journal.pone.0078293

Hébert-Dufresne, L., Scarpino, S. V., and Young, J-G. (2020). Macroscopic Patterns of Interacting Contagions Are Indistinguishable from Social Reinforcement. *Nat. Phys.*. doi:10.1038/s41567-020-0791-20791-210.1038/s41567-020-0791-2

Hohenhaus, P. (1996). *Ad-Hoc-Wortbildung. Terminologie, Typologie Und Theorie Kreativer Wortbildung Im Englischen* Franfurt a. M.: Lang.

Hohenhaus, P. (2006). 'Bouncebackability. A Web-As-Corpus-Based Study of a New Formation, its Interpretation, Generalization/Spread and Subsequent Decline'. *SKASE J. Theor. Linguistics* 3, 17–27.

Huberman, B. A., Romero, D. M., and Wu, F. (2008). *Social Networks that Matter: Twitter under the Microscope*. Available at: http://arxiv.org/abs/0812.1045.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9 (6), e98679. doi:10.1371/journal.pone.0098679

Kerremans, D. (2015). *A Web of New Words*. Bern, Schweiz: Peter Lang. doi:10.3726/978-3-653-04788-2

Kerremans, D., Stegmayr, S., and Schmid, H. J. (2012). "The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change," in *Current Methods in Historical Semantics* (Berlin: Mouton de Gruyter), 59–96.

Kerremans, D., Prokić, J., Würschinger, Q., and Schmid, H.-J. (2019). Using Data-Mining to Identify and Study Patterns in Lexical Innovation on the Web. *Pragmatics Cogn.* 25 (1), 174–200. doi:10.1075/pc.00006.ker

Labov, W. (2007). Transmission and Diffusion. *Language* 83 (2), 344–387. doi:10.1353/lan.2007.0082

Lemnitzer, L. (2010). Wortwarte. Available at: http://www.wortwarte.de/ Accessed 6 August, 2021.

Lu, F. S., Hou, S., Baltrusaitis, K., Shah, M., Leskovec, J., Sosic, R., et al. (2018). Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. *JMIR Public Health Surveill.* 4, e4. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5780615/. doi:10.2196/publichealth.8950

Milroy, J. (1992). *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.

Milroy, J., and Milroy, L. (1985). Linguistic Change, Social Network and Speaker Innovation. *J. Ling.* 21 (2), 339–384. Available at: https://www.cambridge.org/core/article/linguistic-change-social-network-and-speaker-innovation1/EB30A7117CC09F6EDA5255BF9D788D5A. doi:10.1017/s0022226700010306

Nevalainen, T. (2015). Descriptive Adequacy of the S-Curve Model in Diachronic Studies of Language Change. Studies in Variation, Contacts and Change in English 16. Available at: https://varieng.helsinki.fi/series/volumes/16/nevalainen/ Accessed August 6, 2021.

Nini, A., Corradini, C., Guo, D., and Grieve, J. (2017). The Application of Growth Curve Modeling for the Analysis of Diachronic Corpora. *Lang. Dyn. Change.* 7 (1), 102–125. doi:10.1163/22105832-00701001

Pedroche, F., Moreno, F., González, A., and Valencia, A. (2013). Leadership Groups on Social Network Sites Based on Personalized PageRank. *Math. Comp. Model.* 57 (7pp), 1891–1896. doi:10.1016/j.mcm.2011.12.026

Pew Research Center(2019). National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets. Available at: https://www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adultsproduce-majority-of-tweets/ Accessed August 6, 2021.

R Core Team (2018). *R: A Language and Environment for Statistical Computing. Manual*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.Rproject.org/.

Renouf, A., Kehoe, A., and Banerjee, J. (2007). WebCorp: An Integrated System for Web Text Search. Editors M. Hundt, N. Nesselhauf, and C. Biewer. *Corpus Linguistics and the Web.* Amsterdam, New York: Rodopi, 59:47.

Rogers, E. M. (1962). *Diffusion of Innovations*. New York: Free Press of Glencoe.

Schmid, H-J. (2016). *English Morphology and Word-Formation - an Introduction*. 2nd ed.. Berlin: Erich Schmidt Verlag.

Schmid, H-J. (2020). *The Dynamics of the Linguistic System. - Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Spearman, C. (1961). "The Proof and Measurement of Association between Two Things," in *Studies in Individual Differences: The Search for Intelligence* (East Norwalk, CT, US: Appleton-Century-Crofts), 45–58. doi:10.1037/11491-005

Stefanowitsch, A., and Flach, A. (2017). 'The Corpus-Based Perspective on Entrenchment'. In: *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Editors H. J. Schmid. Boston, USA: American Psychology Association and de Gruyter Mouton, 101–127. doi:10.1037/15969-006

Stewart, I., and Jacob, E. (2018). Making 'Fetch' Happen: The Influence of Social and Linguistic Context on Nonstandard Word Growth and Decline. Available at: http://arxiv.org/abs/1709.00345 Accessed August 6, 2021.

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton and Oxford: University Press.

Wang, R., Zhang, W., Deng, H., Wang, N., Miao, Q., and Zhao, X. (2013). 'Discover Community Leader in Social Network with PageRank'. In: *Advances in Swarm Intelligence*. Editors Y. Tan, Y. Shi, and H. Mo. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 154–162. doi:10.1007/978-3-642-38715-9_19

Weinreich, Uriel., Labov, W., and Herzog, M. (1968). 'Empirical Foundations for a Theory of Language Change'. In: *Directions for Historical Linguistics*. Editors W. P. Lehmann and Y. Malkiel. Austin: University of Texas Press Austin, 95–188.

West, Robert., Paskov, H. S., Leskovec, J., and Potts, C. (2014). Exploiting Social Network Structure for Person-To-Person Sentiment Analysis. Available at: http://arxiv.org/abs/1409.2450 Accessed August 6, 2021. doi:10.1162/tacl_a_00184

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *Joss* 4 (43), 1686. doi:10.21105/joss.01686

Würschinger, Q., Elahi, M. F., Zhekova, D., and Schmid, H.-J. (2016). "Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The Case of 'rapefugee', 'rapeugee', and 'rapugee'," in Proceedings of the 10th Web as Corpus Workshop, August, 2016, Berlin, Germany (Berlin: Association for Computational Linguistics), 35–43. Available at: http://aclweb.org/anthology/W16-2605. doi:10.18653/v1/W16-2605

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 27 January 2023
DOI 10.3389/frai.2022.683104

Check for updates

OPEN ACCESS

EDITED BY
Jack Grieve,
University of Birmingham,
United Kingdom

REVIEWED BY
Kevin Tang,
Heinrich Heine University of
Düsseldorf, Germany
Alexis C. Davis,
University of Florida, United States, in
collaboration with reviewer KT
Jon Forrest,
University of Georgia, United States

*CORRESPONDENCE
Taylor Jones
✉ tjones@culturepoint.org

SPECIALTY SECTION
This article was submitted to
Language and Computation,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 19 March 2021
ACCEPTED 09 December 2022
PUBLISHED 27 January 2023

CITATION
Jones T (2023) African American
English intensifier *dennamug*: Using
twitter to investigate syntactic change
in low-frequency forms.
Front. Artif. Intell. 5:683104.
doi: 10.3389/frai.2022.683104

African American English intensifier *dennamug*: Using twitter to investigate syntactic change in low-frequency forms

Taylor Jones*

CulturePoint, LLC., Prince Frederick, New York, NY, United States

There are some linguistic forms that may be known to both speakers and linguists, but that occur naturally with such low frequency that traditional sociolinguistic methods do not allow for study. This study investigates one such phenomenon: the grammatical reanalysis of an intensifier in some forms of African American English—from a full phrase *[than a mother(fucker)]* to lexical word (represented here as *dennamug*)—using data gathered from twitter. This paper investigates the relationship between apparent lexicalization and deletion of the comparative morpheme on the preceding adjective. While state-of-the-art traditional corpora contain so few tokens they can be counted on one hand, twitter yields almost 300,000 tokens over a 10 year sample period. This paper uses web scraping of Twitter to gather all plausible orthographic representations of the intensifier, and uses logistic regression to analyze the extent to which markers of lexicalization and reanalysis are associated with a corresponding shift from comparative to bare morphology on the adjective the intensifier modifies, finding that, indeed, degree of apparent lexicalization is strongly associated with bare morphology, suggesting ongoing lexicalization and subsequent reanalysis at the phrase level. This digital approach reveals ongoing grammatical change, with the new intensifier associated with bare, note comparative, adjectives, and that there is seemingly stable variation correlated with the degree to which the intensifier has lexicalized. Orthographic representations of African American English on social media are shown to be a locus of identity construction and grammatical change.

KEYWORDS

African American English (AAE), lexicalization, social media, language variation and change, morphology, phonology

1. Introduction

Traditional approaches to quantitative sociolinguistics rely on careful elicitation of naturalistic speech with the goal of counting how many tokens of a particular variant a given speaker uses in a given situation, and relating those to both language internal (structural) constraints and language external (social) constraints on the occurrence of a variant. The investigator may use reading passages, carefully constructed interviews with prompts designed to excite the speaker and lower their self-inhibition [in a Labovian framework, the “sociolinguistic monitor” (Labov et al., 2011); in a psycholinguistics framework, introducing cognitive and emotional interference], or may carefully choose questions to elicit data in a rapid, anonymous survey (“where in this store can I find men’s shoes?”). These methods are most effective with tokens that naturally occur with high frequency or that can be easily elicited, for instance deletion (or retention) of postvocalic /r/, or realization of word final ING as either [n] or [ŋ]. However, there are some forms that may be known to both speakers and linguists but which are difficult to elicit and naturally occur so infrequently that traditional sociolinguistic methods do not allow for their study. This is particularly true for African American English, which until recently has primarily been studied by linguists who do not natively speak the variety and who are ethnocultural outsiders (Friedman and Reed, 2020; Hudley et al., 2020), and there is ample evidence that such outsider status can, but does not always, affect data collection in the form of an "interviewer effect" (Rickford and McNair-Knox, 1994; Cukor-Avila and Bailey, 2001). While some features, such as habitual *be* or postvocalic /r/ deletion have been extensively studied, there are other features known to speakers that have received scant or *no* attention in the academic literature (Lanehart, p.c., Smith, p.c., Hall, p.c.). Examples include the associative plural *'nem* (Mufwene, 1998) and the broader change of initial /ð/ to [n] in some phonological contexts, *talkin' 'bout* as a verb of quotation (Cukor-Avila, 2001; Jones, 2016a; Labov, 2018), syntactic change in use of *nigga* (Grieser, 2019; Jones and Hall, 2019; Smith, 2019), and dismissive *bye* among others.

Social media, however, can capture low frequency data that traditional corpora cannot; tokens of interest that may occur a handful of times in a traditional sociolinguistic corpus (e.g., seven instances of third person quotative *talkin' 'bout* and 23 tokens of associative *'nem* in the Corpus of Regional African American Language, Kendall and Farrington 2020) occur hundreds of thousands of times on social media (Jones, 2015). The format is inherently informal (Han and Baldwin, 2011; van Halteren and Oostdijk, 2012; Eisenstein, 2013b), people write for their social networks (Eisenstein, 2013a; Doyle, 2014; Eisenstein et al., 2014; Yuan et al., 2016), and unconventional spellings that pose challenges for traditional NLP applications nevertheless provide rich linguistic information as people engage in identity construction—often through intentionally representing their accents and pronunciation through innovative orthography (Jones, 2016c). People also navigate linguistic taboos orthographically: as Smith (2019) notes, "most white Facebookers (and a few blacks) variably spelled nigga as n***a, nga, ninja, nucca, and nicca, betraying some degree of awareness of the word's taboo status in wider social circles." The usefulness of social media data for investigating low-frequency forms, especially lexical items, is well established (see, e.g., Grieve et al., 2017, 2018). One largely unexplored avenue of linguistic investigation, however, pursued here, is the use of social media as a window into rebracketing, reanalysis, and syntactic change (Eisenstein, 2015; Jones, 2015; Bleaman, 2020; Jones, 2016a,b,c; Austen, 2017; Jones and Hall, 2019).

The object of study of this paper is the previously undescribed syntactic change, from the complement clause "than a mother(fucker)" to the individual lexical item generally pronounced [dɪnəmʌː] (rendered here as *dennamug*) in a vernacular register of African American English. I will refer to this as "intensifier *dennamug*" in what follows. This shift is frequently accompanied by *absence* of comparative morphology: a grammatical shift that is indicative of ongoing reanalysis beyond just phonetic reduction, and which is the focus of this paper (1):

(1) a. It's cold-*er* than a motherfucker
b. It's cold-Ø *dennamug*

There is a small number of counterintuitive exceptions to this generalization, discussed in further detail in Section 4 below (2).

(2) Them 8's and Barkley's availabl**er** dennamug...

Intensifier *dennamug* is rare compared to other lexical items, unstudied, and provides a window into linguistic variation and change in AAE outside of the well-described domain of tense, aspect, and mood. Given the recency of study of AAE, and the focus in sociolinguistics on a handful of topics within the study of AAE (describing the tense/aspect/mood system, status with regards to the creole continuum, the relationship between AAE phonology and literacy in the standard language), not much is known about linguistic variation and change in contemporary AAE as relates to lexicalization, reanalysis, and change at the intersection of phonetics, phonology, and syntax.

Intensifier *dennamug* is evidently the result of a number of different, interrelated linguistic processes: it is an intensifier phrase combined with taboo avoidance, understudied AAE phonology, and competing solutions to the problem posed by phonetic ambiguity. Writing on social media requires authors to derive solutions to the orthography problem posed by standard English orthography's inability to capture some aspects of AAE phonology, and this provides a potential window into the etymological transparency (or opacity) of the intensifier.

The confluence of factors prior to writing is the result of phonetic ambiguity that feeds phonetic reanalysis, which in turn feeds syntactic ambiguity that feeds syntactic reanalysis. In other words, the phrase *than a muh(fucka)* is the origin of a single lexical item, with multiple phonological representations in the speech community, that has very different syntactic properties than its origin—the starting point is a full comparative phrase, and the most advanced syntactic change is a lexical item that modifies a bare (i.e., not comparative) adjective.[1]

The present paper seeks to understand the pathway, and more importantly, the *degree* of lexicalization and reanalysis of *dennamug* using all instances of the term on twitter in a roughly 10 year period that are consistent with reasonable orthographic

[1] There is limited evidence, discussed in Section 4, that this lexical item is on its way toward modifying other types of phrases, as well.

representations of AAE phonology. Here, lexicalization is the degree to which *than a mother* has been from a phrase to a lexical item (e.g., *dennamug*), and reanalysis describes the extent to which it is now treated as a lexical intensifier (which therefore no longer requires comparative morphology on the adjective it modifies). The first sections of the paper describe the phenomenon under investigation, and the subsequent sections investigate the degree of lexicalization and reanalysis using quantitative methods drawing on a corpus of tweets specifically gathered to investigate this topic. Twitter is a mechanism, albeit an imperfect one, for the study of *dennamug* because, despite its low frequency in conversation and the difficulty eliciting it in a traditional sociolinguistic interview setting, there are hundreds of thousands of tokens, and the written format forces speakers to choose whether they write the intensifier as a single word or as a phrase, what the phonological components of the intensifier are, and whether it is accompanied by comparative morphology. Moreover, traditional sociolinguistic corpora are not viable for the present study, as there are no tokens of *dennamug* present in the Corpus of Regional African American Language (CORAAL, Kendall and Farrington, 2020) or Corpus of Contemporary American English (COCA, Davies, 2008), to take two well-respected examples, and only handful of tokens of *than a motherfucker* in CORAAL and 36 such tokens in COCA.[2] Moreover, in writing, there is no "in between," as there is in fast casual speech—authors are forced to make choices about how to represent their language that do not allow for ambiguity. The process of reanalysis resulting in intensifier *dennamug* is therefore a perfect illustration of the value of novel computational approaches to sociolinguistics, using social media data, in an area where traditional sociolinguistic methods fail.

Before discussing the materials and methods, it is necessary to discuss lexicalization, intensifiers, comparative phrases, AAE phonology, and taboo avoidance, and to further describe the phenomenon under investigation, as despite the fact that its use is widespread it is nevertheless previously unattested in the academic literature on African American English.[3] I will treat these in turn in the following sections. With this foundation, I can then return to materials and methods for the present study, which focuses on the extent to which a semantic shift has occurred following fusion and coalescence (Section 1.1), as evidenced by a change in obligatory morphological marking on the adjective *than a mother* ∼ *dennamug* modifies.

1.1. Lexicalization

Lexicalization, following Brinton and Traugott (2005), is "the process by which new items that are considered 'lexical'...come into being." Lexicalization is often contrasted with "grammaticalization," which refers both to a linguistic phenomenon and field of study (Hopper and Traugott, 2003). The field occupies itself with the "part of the study of language change that is concerned with [...] how lexical items and constructions come [...] to serve grammatical functions or how grammatical items develop new grammatical functions," and within the field the "steps whereby particular items become more grammatical through time" is referred to as *grammaticalization* (Hopper and Traugott, 2003, pp. 1–2). The distinction between lexicalization and grammaticalization is not always clear, especially in the domain relevant to *dennamug*, in which processes of fusion result in decreased compositionality (Brinton and Traugott, 2005). Indeed, examples of fusion and coalescence, to be defined below, have been treated either as lexicalization or grammaticalization by various researchers, including phrases that have become fixed (e.g., *today* < OE to + dæge "at day-DAT"), derivational affixes derived from roots in compounds, some fixed phrases, multiword verbs, composite predicates or complex verbs (e.g., "lose sight of," "take action," "make use of"), and phrase discourse markers (e.g., "I mean") (Brinton and Traugott, 2005, p. 63–67). Following Wischer (2000), the present study treats the development of *dennamug* as an instance of lexicalization, rather than grammaticalization, because as boundaries and syntactic structure are lost, a specific semantic component is added, rather than semantic components being lost with categorical or operational meaning foregrounded (Wischer, 2000, p. 364–365).

While a broad range of phenomena contribute to lexical innovation, including compounding, derivation, conversion clipping, blending, back formation, and initialisms, among others, the most relevant aspects of lexicalization to the present study are those that relate to reanalysis and change over time, namely univerbation, demorphologization, and idiomaticization. Univerbation is the "unification…of a syntactic phrase or construction into a single word" (Brinton and Traugott, 2005, p. 48–51). A subset of univerbation, sometimes called "delocutivity" (Benveniste, 1971), obtains when an entire phrase is transformed into "a more or less complex word expressing a contiguous concept," (Blank, 2008, p. 1602, 1604), as in Italian *non so che* "I don't know what" > *nonsoche* "something that is difficult to explain" Spanish *vuestra merced* "your honour" > *usted* "you (formal)" and English *goodbye* from *God be with you*. Some argue that while rare, these are exemplars of lexicalization because they are not

[2] Some of these tokens are from non-AAE speakers writing for AAE-speaking characters, as in *The Wire*, others are written and delivered by AAE speakers as in *Friday*, but COCA has the written script, and not what was actually spoken.

[3] This is another reason for using quantitative methods: linguists unfamiliar with the term may doubt its existence; the data for the present analysis and the analysis itself are independently verifiable and do not rely on readers trusting an analysis of an unverifiable data set, as would be the case in traditional ethnographic and sociolinguistic field work.

just fusion—the obligatory collocation of previously separable material—but also of conversion, in which an item shifts from one category to another. In the case of *dennamug*, the most extreme forms are fully univerbated, and have gone from a comparative phrase to a lexical adverb (Blank, 2008).

Demorphologization describes a process "whereby a morpheme loses (most of) its grammatical-semantic contribution to the word and becomes an indistinguishable part of the construction of the word, while retaining part of its original phonological substance." (Brinton and Traugott, 2005, p. 52). Indeed, we find demorphologization in *dennamug*, as AAE speakers who use it, and authors in these data, are frequently unaware of any connection to mother and disagree about whether the last syllable is *mub*, *mud*, or *mug* (see example 6 below).

Lastly, idiomaticization is the extent to which a construction is more idiom-like. What exactly this means in practice is a matter of debate, however Brinton and Traugott (2005) characterize it as comprising three components:

1) **Semantic opacity or noncompositionality**: it is impossible to deduce the meaning of *shoot the breeze* from "shoot" + "the" + "breeze."
2) **Grammatical deficiency**: an idiom does not permit the syntactic variability characteristic of free combinations such as passive(**the breeze was shot*) negation (*?didn't shoot the breeze*), internal modification (*shoot a strong breeze*, **shoot breezes*, **shoot some breeze*), or topicalization (**the breeze he shot*).
3) **Lack of substitutability**: synonymous lexical items cannot be substituted (**shoot the wind*, **fire at the breeze*), nor can items be reversed or deleted.

Relevant intensifiers in informal AAE span a full spectrum between clearly non-idiomaitic, compositional and substitutable (*than a mother* ∼ *than a bitch*), through moderate univerbation and demorphologization while still exhibiting some level of compositionality and substitutability (*danna muv* ∼ *danna bish*), to fully univerbated, demorphologized, and idiomaticized (*dennamug*). The best evidence for reanalysis is not just univerbation, demorphologization, and idiomaticization, but subsequent changes elsewhere in the clause or sentence. As will be shown below, the greater degree to which *dennamug* exhibits these characteristics of lexicalization, the greater the likelihood that the adjective *dennamug* modifies appears as a bare adjective, without comparative morphology, because *dennamug* is functioning differently grammatically than the comparative phrase *than a mother*. Before demonstrating this, it is necessary to discuss the morphosyntax of intensifiers (Section 1.2) comparative phrases (Section 1.3) and the interactions between relevant AAE phonology (Section 1.4) and taboo avoidance (Section 1.5).

1.2. Intensifiers

Intensifiers are words or phrases that do not modify the propositional meaning of a clause, but add force. They are more or less semantically vacuous, although the degree to which they are more or less depends on the intensifier and context. Intensifiers modifying adjectives come in two types, depending on the adjectives they modify: attributive and predicative (Tagliamonte, 2012).

Attributive intensifiers modify attributive adjectives and can precede or follow the adjective they modify (3):

(3) a. a cold *ass* day
  b. *really* cold day

Predicative intensifiers modify predicate adjectives:

(4) a. she's *so* fine
  b. she's *really* sweet (Wilson and Gordie, 1957)

Some intensifiers, like *really*, can serve as both attributive and predicate intensifiers, some, like *ass* can only serve as attributive intensifiers, and some, like *deadass* can only serve as predicate intensifiers. Note that *-ass* originates in African American English (Spears, 1998; Collins et al., 2008; Miller, 2017), and that *deadass* is the result of a number of steps of reanalysis: *I'm serious* > *I'm dead serious* > *I'm dead ass serious* > *I'm deadass* > *deadass* + adjective (e.g., *I'm deadass hungry*) > *deadass* + predicate (e.g., "Cuomo is deadass trying to kill us all" apropos of restaurant reopenings in New York during the COVID-19 pandemic).[4] Note also that *deadass* is the result of grammatical reanalysis of an earlier form, itself the result of reanalysis.

Some intensifiers have historically come from comparisons that lose semantic force, for instance, *pitch* "extremely." Originally a comparison referring to the black resin used to caulk sailing vessels called *pitch*, as in *as black as pitch* or *pitch black*, *pitch* now modifies other verbs of perception, as in *pitch quiet* and *pitch silent* (5).

(5) Our old neighborhood was perfect. It was pitch quiet.[5]

A similar case is *as hell*, which no longer draws a comparison to a specific conception of the afterlife, as in *pleasant as hell* "very pleasant."

The object of study in the present paper is a predicative intensifier (although more will be said about this

[4] https://twitter.com/QueeringPsych/status/1355329178695041025

[5] https://twitter.com/DomoDash_/status/1344495542882201603

characterization in Section 4), derived from a comparative phrase. In the next section I discuss comparatives, as reduction of comparative morphology is one form of evidence that *dennamug* is the result of lexicalization and grammatical reanalysis.

1.3. Comparatives

Comparatives are similarly complex, and vary across multiple parameters including whether they are bound or periphrastic, clausal or phrasal, and whether they express equality or inequality of degree. English has both bound and periphrastic comparatives (also called synthetic and analytic), hypothesized to be sensitive to the number of syllables the adjective comprises, so bound comparatives are preferred for monosyllables (*easy* + *-er*), and adjectives with more than two syllables almost always occur with a periphrastic comparative (e.g., *more intelligent*, compare to *intelligenter* Jespersen, 1949; Cygan, 1975; Bauer, 1994; Leech and Culpepper, 1997; Lindquist, 2000; Enzinna, 2017). Clausal comparatives take a clausal complement (*Mary is taller than Susan*) whereas phrasal comparatives do not require a comparative clause and may instead use case marking, for instance, the adessive case in Hungarian (Backsai-Atkari, 2014, p. 4). The latter does not occur in English, and therefore will not be discussed further here. Comparatives can express equality (*He's **as** dumb **as** a brick*) or inequality (*He's dumb**er than** a brick*, *He's **less** intelligent **than** a bag of hair*).

The exact syntactic structure of comparatives has been a matter of lively debate since at least the 1970s, with various structures proposed by Bresnan (1973), Izvorski (1995), Corver (1997), Lechner (1999), Lechner (2004) and Backsai-Atkari (2014), among others. The latter, relying on an analysis that makes use of both Quantifier and Degree Phrases, is assumed here (Figure 1). This is important, because grammaticalization is the result of both rebracketing and reanalysis. Not only does this reanalysis of *than a mother* dramatically change the assumed syntactic categories of its component parts, it also results in an unusual constituent order: an adverb following the adjective it modifies. The important point to note about Backsai-Atkari's proposed structure is that the comparatives *-er* and *more*, while occuring in different syntactic positions, are performing the same function, and that what follows is a clausal complement. The rebracketing and reanalysis of *than a mo(ther)* > *dennamug* is made easier by the fact that AAE phonology is not well served by standard English spelling conventions, and that many expressions and words in AAE are not described in any style guide or dictionary, leaving it to the speakers themselves to determine how to map sounds to spelling. The next section discusses relevant AAE phonology and the following discusses taboo avoidance and deformation relevant to the phonological shape of *than a mo(ther)* as it lexicalizes.

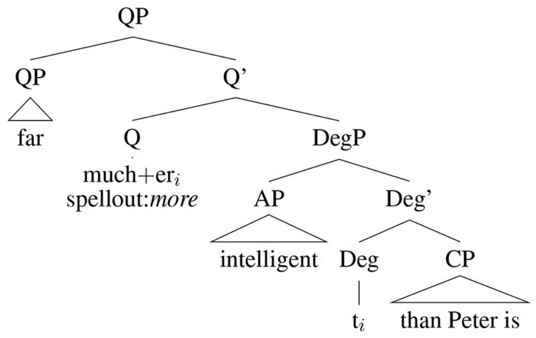


FIGURE 1
Comparative structure after Backsai-Atkari.

1.4. Relevant AAE phonology

An enormous amount has been written about the phonology of AAE, although the focus of much of the sociolinguistic inquiry into AAE has been a relatively small handful of phonological features. Erik Thomas and Guy Bailey summarize the broad strokes in their 1998 and 2015 papers on the subject (Bailey and Thomas, 1998; Thomas and Bailey, 2015), however regional variation in AAE phonology is understudied, and there are understudied phonological features relevant here. Well known variables associated with AAE inlcude variation in -ING, postvocalic /r/ vocalization and deletion, postvocalic /l/ vocalization and deletion, and so called TH-STOPPING and TH-FRONTING. Less known and under-researched phonological variables include stop devoicing, debuccalization, and deletion (Farrington, 2018), vowel nasalization, and postocalic /v/ deletion (mentioned in Thomas, 2007 and Jones, 2016b).

There are multiple possible pathways from *mother* to [mʌː] in AAE, and indeed we find that other words are subject to the same processes (cf. *brother* > *brer* ∼ *bruh* [brʌː]). The word *motherfucker* itself was subject to grammaticalization: it has undergone semantic broadening from an epithet suggesting taboo sexual relations to an individual lexeme that serves as an purpose exclamation, and it is in this context that the first word has undergone reduction. It is often rendered *muhfugga* in writing on social media, reflecting actual pronunciation, attested as early as 1995[6]:

> Smokey: "You know what they say, the older the berry, the sweeter the juice."
> Craig: "n—, it's the *blacker* the berry."

[6] Note that there is a range of pronunciations of the last two syllables as well, from [fʌkə] to [fʌgə] to [fʌː].

> Smokey: “Yeah, well, she blacker than a motherfucker [dɛn ə ˈmʌfəkə], too.”
>
> (Gray, 1995)

Perhaps interestingly, the earlier widely available recordings of the word *motherfucker* in AAE, as in Richard Pryor’s stand up comedy from 20 years prior, are also reduced, but not to the same extent. They also co-occur with a possibly bare adjectives:

> I don’t blame ’em. Be in a cave two thousand year that’ll make you *mad(der) than a motherfucker* [mæd$^{(ə)}$ nə mʌːv.fʌkə], won’t it?
>
> (Richard Pryor, “Mudbone Goes to Hollywood” at the Pryor, 1976)

The process that changes the initial phoneme in words like “than” and “them” to [n], and postvocalic /r/ deletion, along with prosodic factors mean that it is difficult to state with certainty whether Pryor’s grammar includes bare adjectives in comparative constructions (and a short epenthetic schwa appears because nasal plosion is not an option in a /dn/ sequence), without further research.

Possible pathways of phonological change leading to [mʌː] include:

1. /r/ vocalization and deletion > TH-STOPPING > elision of schwa > postvocalic stop deletion
2. /r/ vocalization and deletion > TH-FRONTING > elision of schwa > postvocalic /v/ deletion
3. /r/ vocalization and deletion > TH-FRONTING > elision of schwa > voicing assimilation on postvocalic /v/ preceding onset /f/

Regardless the specific phonological pathway, the result is that *motherfucker* is pronounced, and frequently written, as *muhfucka* or *muhfugga*, and it is this pronunciation that is the starting point for *dennamug*.

The crucial factor here is that the surface phonology of some varieties of AAE allow lax vowels in open syllables, so seemingly un-checked wedge occurs in words like [mʌː] “mother,” [bɹʌː] “brother,” and [lʌː] “love,” but that these are instances of what Farrington (2018) calls *incomplete neutralization* (in his case, discussing apparent deletion of word final coronal stops, whose voicing specification are still recoverable by vowel length). In this case, a closing consonant is implied, and people writing on social media are compelled to choose one of either <d>, <g>, <v>, <b>, or the generic <h> to avoid readers imagining the unwanted pronunciation /mu/. Some of these forms are more suggestive of truncation (<muh>) or regional AAE phonological processes (<muv>) and others are more phonologically opaque (<mub>, <mud>, <mug>). If speakers do not hear or produce a word final consonant, but know that the surface string has a phonologically illicit long lax vowel, then they can infer that there is a closing voiced consonant, but may be unsure what the precise nature of the consonant is. Perhaps unsurprisingly, there are metalinguistic discussions about which spelling of *dennamug* is “correct”:

(6) a. Dennamug or Dennamud ?[7]
b. Das no typo tho. “than a mub” is a southern saying.[8]
c. yes it is . It a typo cuz its *mug (in response to 6b).[9]

Surprisingly, the responses all indicate ...<mug> is “correct” and none claim *than a mother* is technically correct. While it is also theoretically possible that the /f/ in /mʌfʌgə/ underwent lenition, and the final schwa underwent apocope, resulting in the change /mʌfʌgə/ → mʌːgə → mʌːg → mʌː as a third possible pathway, there is no literature on AAE phonology that would support such a change (i.e., the deletion of an intervocalic fricative when the vowels on either side are the same), and no comparable examples, to my knowledge.[10] Instead, the last element, to which we now turn, is taboo avoidance.

1.5. Taboo avoidance

Taboo avoidance is cross lingusitically common and takes many forms. Different languages may treat different classes of words as taboo, so for instance *ukuhlonipa* “politeness” in the Nguni languages requires deformation of phonemes or syllables related to the names of family by marriage, resulting in a rich set of synonyms. In most varieties of English, the words subject to taboo avoidance are scatological, sexual, and religious in nature (Allan and Burridge, 2006). One form of avoidance is taboo deformation, which can take many forms: minced oaths (*god damn it* > *gosh darn it*), rhyme (*bloody* > *ruddy*), metrical substitutions (*shut the fuck up* > *shut the front door*, or *motherfucker!* > *mother father!*), deletion, and acronyms (*as fuck* > *A. F.* > *ayeff*).

Motherfucker is a taboo word, even thought it may be somewhat meliorated in some informal varieties of AAE (see Spears 1998 and Jones and Hall 2019 for other examples of “so-called obscenity”). While there are multiple strategies for avoiding the taboo, such as the metrical substitution *mother father!*, the strategy relevant here is simply beginning the taboo word, and not finishing it. That is, deletion of most of the phonological material: *mother~~fucker~~*. As an expression like *hungrier than a mother* is itself subjected to this treadmill, the result it *than a mo~~ther~~*.

[7] https://twitter.com/bydie_ching2x/status/547810865442226176

[8] https://twitter.com/MoeMyGod/status/229455319460356096

[9] https://twitter.com/JaCorii_/status/229457265239928833

[10] For instance, in varieties of AAE that front /ð/, as in mʌvə “mother,” there is no attested evidence of such reduction, as in *my mʌː “my mother.”

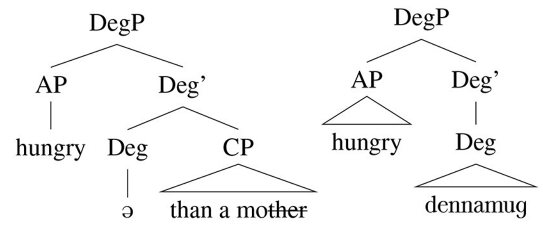


FIGURE 2
Syntactic reanalysis.

As discussed in Section 1.1, one view of lexicalization relies on reanalysis, and by this view true lexicalization requires rebracketing and a change in the assumed hierarchical structure of a phrase. The combination of a phonological reduction of *mother* in *motherfucker* ([ˈmʌːfʌɡə]) and taboo avoidance creates the perfect conditions for reanalysis, where the syntactically complex [dɪn [ə [mʌː]]] is reanalyzed as a single word (Figure 2).

Furthermore, widespread postvocalic /r/ deletion in AAE means that the comparative morpheme *-er* may be realized as a schwa only, and one or two syllable adjectives may contribute to a prosodic pressure toward further reducing the schwa, especially when the following intensifier is not strictly recoverable as a comparative phrase, and especially in fast speech (see, e.g., Davidson, 2006 on pretonic schwa deletion, and compare against AAE *bednot* “better not”):

(7) a. $\overline{\text{fat}}$ ĕr $\overline{\text{than}}$ ă $\overline{\text{mug}}$

b. $\overline{\text{hap}}$ pў ĕr $\overline{\text{than}}$ ă $\overline{\text{mug}}$

In the next section, evidence for metalinguistic discussion of *dennamug* among AAE speakers is adduced as further evidence for a cline of lexicalizing forms. In Section 2, I turn to the materials and methods for the present study, investigating the extent to which there is quantitative evidence for rebracketing and reanalysis.

1.6. Metalinguistic awareness as evidence for reanalysis

Further evidence for possible lexicalization comes from how speakers themselves discuss the language. Beyond exchanges on twitter like that in example 6, above, there is evidence that some AAE speakers believe *dennamug* to be a lexical item and not a comparative clause. Urban dictionary has an entry from 2006 for *than-a-mug* with the definition “To the extreme of something’s current state,” and an entry from 2003 for *dennamug* with the definition “hella, a lot, very much.”[11],[12] Absent is any reference to the expression “than a motherfucker.” The latter definition predates the launch of twitter by 3 years, demonstrating that *dennamug* cannot be merely an orthographic meme on social media. Evidence of speaker perceptions are not limited to written attestations, either: for instance, in 2013, YouTube personality Kevin Fredericks (known as Kev On Stage) made a video of “Black Folks Slang” in which explains *dennamug* (which he spells on screen as <dennamug>) is “the measure of unit of something that is something else” and elaborates that “whatever *dennamug* is, that unit measure, you’ve gotta be doing *more* than that.”[13] While this is not necessarily how a linguist would phrase it, it is nevertheless clear that he is describing an intensifier.[14] He pronounces a word final /g/ in citation form, but then provides six example sentences of his own with no closing /g/.[15] At no point does he make reference to the comparative phrase *than a mother*. Of the six examples of his own he provides, five had comparative morphology, and one did not, of the five examples he provides from others, none had comparative morphology.

It should be clear from the above that *dennamug* is (1) a phenomenon in spoken AAE that (2) speakers expect others to understand, and as such, it is widespread (occuring in movies, radio, television, stand up comedy, YouTube videos, get-out-the-vote ads,[16] political rallies with former presidents,[17] etc.), and (3) its origins as a comparative phrase are opaque to some speakers, who now perceive it to be a single word, and who no longer consistently use it with comparative morphology. It is historically related to, but distinct from, the comparative phrase *than a motherfucker*. Despite being widespread, it is just the type of phenomenon that is difficult or impossible to study using traditional sociolinguistic methods and corpora. However, this is precisely a situation in which new computational methods, in this case as simple as web scraping, allow for sociolinguistic insights. Lexicalization is generally understood to be a slow process that unfolds over time, and one for which both older and newer forms overlap. Moreover, reanalysis does not necessarily entail immediate change in surface manifestations (Langacker, 1977), but such change is a strong piece of evidence for reanalysis. In the next section, I discuss materials and methods

[11] https://www.urbandictionary.com/define.php?term=than-a-mug

[12] https://www.urbandictionary.com/define.php?term=dennamug

[13] https://www.youtube.com/watch?v=RRriNPDpmGU

[14] His description implicitly agrees with the last few decades of research on comparative phrases, which situate them in DegP, as well.

[15] This was determined perceptually and with the aid of spectrographic analysis in Praat, which confirmed there is no “velar pinch” consistent with a closing velar consonant.

[16] https://deadline.com/2022/09/cedric-the-entertainer-hershel-walker-video-1235120898/

[17] https://news.yahoo.com/woman-tells-obama-hes-finer-211110070.html

used to investigate the extent to which reanalysis has occurred, using absence of the comparative morpheme *-er*—obligatory in comparative phrases—as an indicator of reanalysis and lexicalization.

2. Materials and methods

For the present study, I gathered all tokens of all spellings of *dennamug* on twitter in the 10 year period from 2007 to 2017 that are consistent with reasonable orthographic representations of AAE phonology. To do so, I accounted for TH-STOPPING by searching for both an initial <th> and initial <d>; I accounted for raising of /æ/ and the PIN-PEN merger by searching for <a>, <e> and <i>; I accounted for one-, two-, and three-word spellings and accompanying duplication of orthographic <n> in single and two-word spellings (as in <dinna muh>); and I searched for word final <d>, <g>, <v>, and <h>. I did not initially search for word final <b>; however, I later gathered all 143 tokens manually. While it is possible there are unaccounted for spellings, they are so rare as to be irrelevant to the analysis here (in fact, many of the tokens generated by this algorithm returned one or no tweets). I used a shell script to generate all possible spellings meeting the above criteria and to make individual calls to the now deprecated `get-tweet` script in Python.[18] The tokens sought were thus any that matched:

```
(th|d)(a|e|i)n+\\s?a\\s?mu(b|d|g|v|h)
```

The resulting data set comprised 294,364 tweets, plus another 143 observations with a word final <b>, for a total of 294,507 tweets. After eliminating false positives (e.g., *she's uglier than a mud fence*, or *nothing better than a mug of hot chocolate*), eliminating tweets that included some spelling of *fucker* immediately following the token, and eliminating false positives in other languages,[19] 264,816 tweets remained.

Examples of true positives include:

- I know terence blanchard bouta be playing that trumpet louder than a mug lol.
- Back sore dan a mug from rehearsal i could use a back rub.
- Man one of my friends is long winded den a mug dawg she can talk yo fuckin ear off.
- Pook auntie funny den a mud.[20]

Data gathered included tweet ID, username, tweet text, date, time, retweets, likes, geolocation (where applicable), mentions, hashtags, and permalink. I created variables for which token was contained in the tweet (e.g., *danna muv*), preceding adjective, and whether comparative morphology was present or absent. User profile pictures were also collected. While, in principle, these could be used to code gender (or more accurately, gender presentation), approximate age, and (not self-identified) "race," those were not coded for in the present analysis and remain an area for future inquiry, however, there is no *a priori* reason to think gender and age are relevant to use of *dennamug* and race is only relevant insofar as it is a highly correlated but imperfect proxy for use of AAE. [21] Nevertheless, visual inspection of the profile picture data suggest that the subjects are not imbalanced by gender, and are overwhelmingly Black and American. Unfortunately, visually inspecting and hand coding for apparent age and gender presentation was unfeasible. Such inspection and coding would also be fraught with methodological and ethical challenges, including but not limited to own-age and own-race biases on the part of the researcher and image misattribution (for instance, when a twitter user's profile picture is of a relative, celebrity, or other person who is not the author). The rest of the language in the tweets exhibits both orthographic representations of AAE phonology (PIN-PEN merger, coda cluster simplification, TH-STOPPING, TH-FRONTING, etc.), AAE morphosyntax (e.g., habitual *be*, stressed *been*, preterite *had*, copula deletion, etc.), and AAE lexical items that have not yet been borrowed by the white mainstream (e.g., *saditty*, *bama*, *ashy*, *jont*, *darkskin*, *geeked*, *siced*, etc.). I normalized the most common variant spellings, changing word final <a> and <ah> to <er> and normalizing arbitrarily many repeated letters, as in <sleeeeepy> to <sleepy>, but did not normalize other respellings that were not merely lengthening, as in <fye> "fire" or <asapidlier> *ASAPedly-er* "quicker."[22]

The two users who tweeted the most, NICKNCEJAIGH with 615 tweets and 101THEGREAT with 478 tweets, were marketing their original songs "Harder than a mug" and "Fresher than a muh," respectively, both of which use the intensifier in the hook. Because the tweet texts, while unique, were using

[18] Twitter has historically been antagonistic to those who would scrape data, and has changed the way tweets are represented in a browser to render this script and others like it ineffective. The primary mechanism was to automate endless browser scrolling (and refreshing) and parse the html.

[19] After removing other character sets, I used automatic language detection with `cld3` to eliminate non-English tweets (mostly Indonesian). I manually checked these, as language detection is unreliable with AAE (Blodgett et al., 2016, 2020). For instance, of over 400 tweets categorized as Spanish, only 11 were actually in Spanish. The rest were evidently characterized as Spanish due to the presence of English discourse marker *yo*, or mention of place names in the Southwest (e.g., El Paso).

[20] "Pook's auntie is extremely funny."

[21] Within a Labovian framework it may be reasonable to hypothesize that if use of *dennamug* isa change in progress, that it may be driven by young women, however these data do not permit us to investigate this possibility. Anecdotally, this author hears *dennamug* in daily life from both genders and across ages.

[22] Derived from the initialism for "As Soon As Possible."

dennamug in citation and not actual use, I did not retain tweets from these two authors. Because they shared links to their songs, however, there is further evidence for pronunciation. Despite the orthographic representation, both dramatically phonetically reduced the intensifier, with NICKNCEJAIGH saying [ʃi goː hɑːdə dɪnəˈmʊːh] "she go harder *dennamug*," and 101THEGREAT singing [ɑ̃ frɹɛʃə dĩnəˈmʌːh] "I'm fresher *dennamug*."[23]

The fifty most common spellings of intensifier *dennamug* are presented in Figure 3.[24] Because these follow the expected Zipfian distribution, they are presented log transformed. The fifty most common adjectives, after spelling normalization, are presented in Figure 4.

Of primary interest was the relationship between orthographic indicators of reanalysis as a single, opaque lexical item, and presence or absence of comparative morphology. To investigate this I performed both traditional logistic regression, and mixed effects logistic regression to account for unmeasured author characteristics. The response variable was presence of comparative morphology on the adjective (that is, some orthographic representation of a final *-er*). The predictor variables were the presence or absence of orthographic representation of TH-STOPPING (initial <th> or <d>); the orthographic representation of initial vowel (with <a> as a reference category); the orthographic representation of the final "closing" consonant (with <g> as a reference category);[25] complexity (meaning, how many orthographic words); lemma frequency (where *hungry*, *hungrier*, and *hongryyyyyy* all count as tokens of HUNGRY); whether the lemma was *good* (to account for the now enregistered *gooder dennamug*); and random intercepts in the mixed effects model for username to account for unmeasured author characteristics. The form of the intensifier used was not included, as the first four variables completely and uniquely describe it (for instance, TH-stopping, an initial /e/, complexity of 1, and final consonant /g/ selects *dennamug*), and any model that included it would suffer from severe multicollinearity. Similarly, adjective lemma was not included, as lemma frequency was highly correlated with it.[26]

Because the number of tweets per author followed a power distribution with the vast majority of user IDs associated with only a single tweet, mixed-effects logistic regression with a random term for username was not feasible on the full data set. However, it is possible that unmeasured author characteristics had an effect on rate of comparative morpheme deletion. To overcome this limitation, I performed logistic regression without a random term for username on the full data set, and performed goodness-of-fit tests, then performed logistic regression on the subset of the data comprising authors who tweeted only once, and performed mixed effects logistics regression with a random term for username on a subset of the data that encompassed all users who tweeted *at least* 10 times,[27] which represented 52,646 observations from 2,442 authors. The results were consistent and robust across multiple specifications of the model.

The distribution of final consonants was heavily skewed (toward <g>), as was the distribution of orthographic complexity (with two spaces heavily preferred, followed by none). Because the final consonants are easily divided into two natural classes (i.e., voiced stops, comprising /b/, /d/, and /g/, and fricatives, comprising /h/ and /v/), the model was run with a binary variable for *fortition*. Similarly, because any reduction in orthographic complexity is a sign of reanalysis, the model was run with a binary variable for *complexity*. This form of the model significantly outperformed others that had five categories for final consonant and three categories for orthographic complexity.

The form of the basic model was therefore:

$$\begin{aligned} Comparative = \beta_0 + \beta_1 THstopping + \beta_2 Vowel \\ + \beta_3 fortition + \beta_4 Complexity \\ + \beta_6 LemmaFrequency + \beta_7 isGood + \epsilon \end{aligned} \tag{1}$$

and for mixed effects logistic regression:

$$\begin{aligned} Comparative_{ij} = \beta_0 + \beta_1 THstopping_{ij} + \beta_2 Vowel_{ij} \\ + \beta_3 fortition_{ij} + \beta_4 Complexity_{ij} \\ + \beta_6 LemmaFrequency_{ij} + \beta_7 isGood_{ij} \\ + \beta_8 Author_j + \epsilon_{ij} \end{aligned} \tag{2}$$

Model comparison and post-estimation tests confirmed that these models outperformed similar models that dropped variables included in these models. It also dramatically outperformed models that included *year*, which was not significant in the models that included it, and in some cases caused failure to converge. Vowels other than <a>, final consonants other than <v> and <h>, and reduced orthographic complexity (suggestive of univerbation), were anticipated to be associated with greater comparative deletion. TH-STOPPING was not expected to be associated with change in comparative morphology, as it is a productive process in AAE phonology. Use of the lemma *good* was expected to be

[23] These transcriptions were confirmed by two phoneticists who do not speak AAE, who were not told what language the audio was, and who heard the utterances in isolation and were therefore not predisposed to a particular analysis.

[24] The reader may note that my preferred spelling, which helps to make the intensifier under discussion maximally distinct from the comparative phrase, is ranked number 10.

[25] While <v> or <h> might make more sense linguistically, <g> was much more common in the data.

[26] *Almost* perfectly, as variation in spelling (e.g., *hongry*, *hungryy* introduced some noise).

[27] I chose 10 or more tweets to ensure that there were sufficient observations to allow for a random term for username.

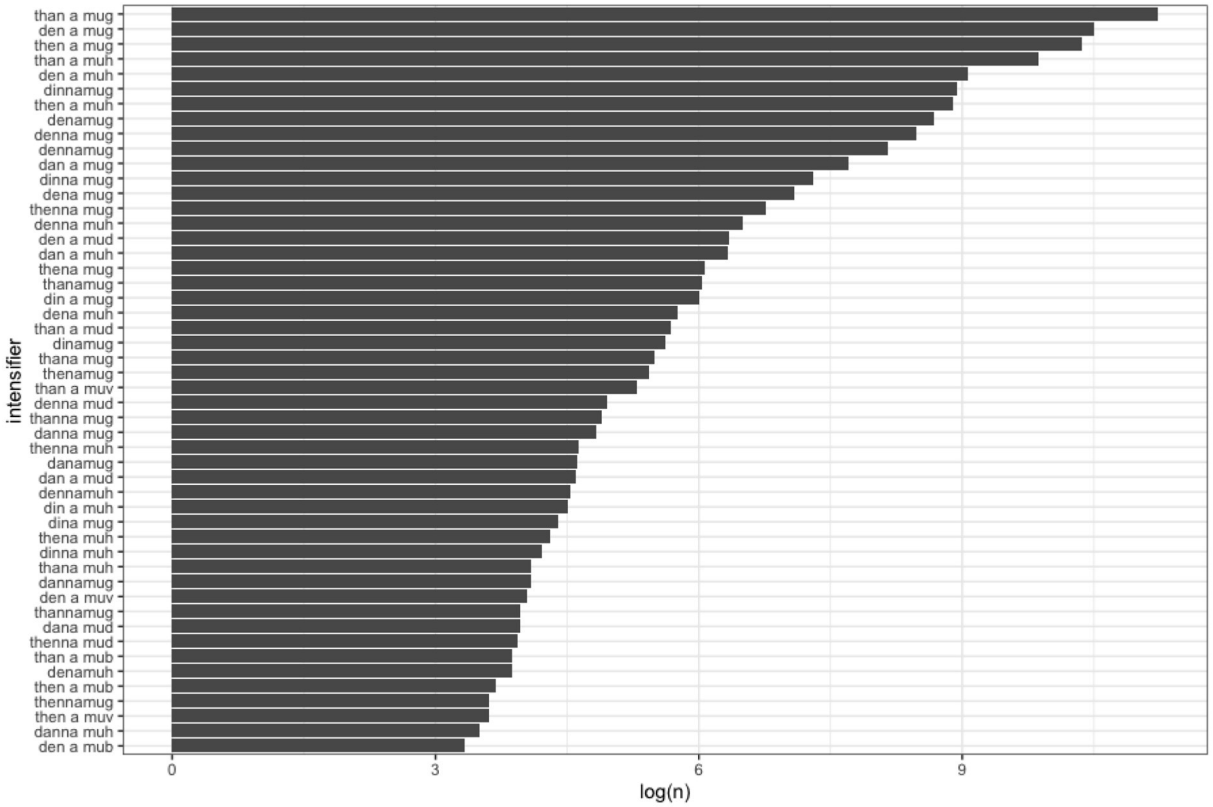


FIGURE 3
Fifty most common intensifiers (log transformed).

associated with an increase in comparative morphology, because of the enregistered idiom *gooder than*.... Lemma frequency was likewise expected to be associated with greater use of comparative morphology.

3. Results

The results of the logistic regression performed on the full data set are presented in Table 1. All predictor variables are significant at the 0.001 level. The intercept of 0.43 indicates that all things being equal, the probability of encountering comparative morphology on the preceding adjective was only 60.4%. TH-STOPPING was associated with a small, but significant *increased* probability of encountering comparative morphology on the preceding adjective (see below for discussion). Initial vowels other than <a> were associated with a significant decrease in probability of comparative morphology (33 percentage points for <e> and 28 percentage points for <i>). Fortition of the final consonant, and reduced orthographic complexity were both associated with significant decreases in probability of encountering comparative morphology. Lemma frequency was associated with a small but positive effect—more common words were more likely to exhibit comparative morphology, all things being equal. The word *good* was associated with a significant, positive effect: if the lemma was *good*, it was much more likely for the form of the word to be *gooder*, regardless of the form of the following intensifier (as anticipated). All things being equal, comparative morphology on the adjective was associated with a probability of 0.15 of appearing before a univerbated intensifier with initial /d/, raised first vowel, and a word-final stop (e.g., <dennamug>, <dinnamud>, etc.). That is, univerbation and phonological opacity obscuring the relationship to *than* and *mother* were associated with a dramatic loss of comparative morphology on the adjective.

Performing logistic regression on the subset of tweets for which the author only tweeted once, the results are similar, and are presented in Table 2. In this subset of the data, TH-STOPPING is no longer a significant predictor. The effect directions are the same, and the magnitudes are approximately the same as in the model on the full data set, except for the effect for orthographic complexity, which is larger by a factor of three.

Finally, the results of mixed-effects logistic regression accounting for unmeasured author characteristics on those who

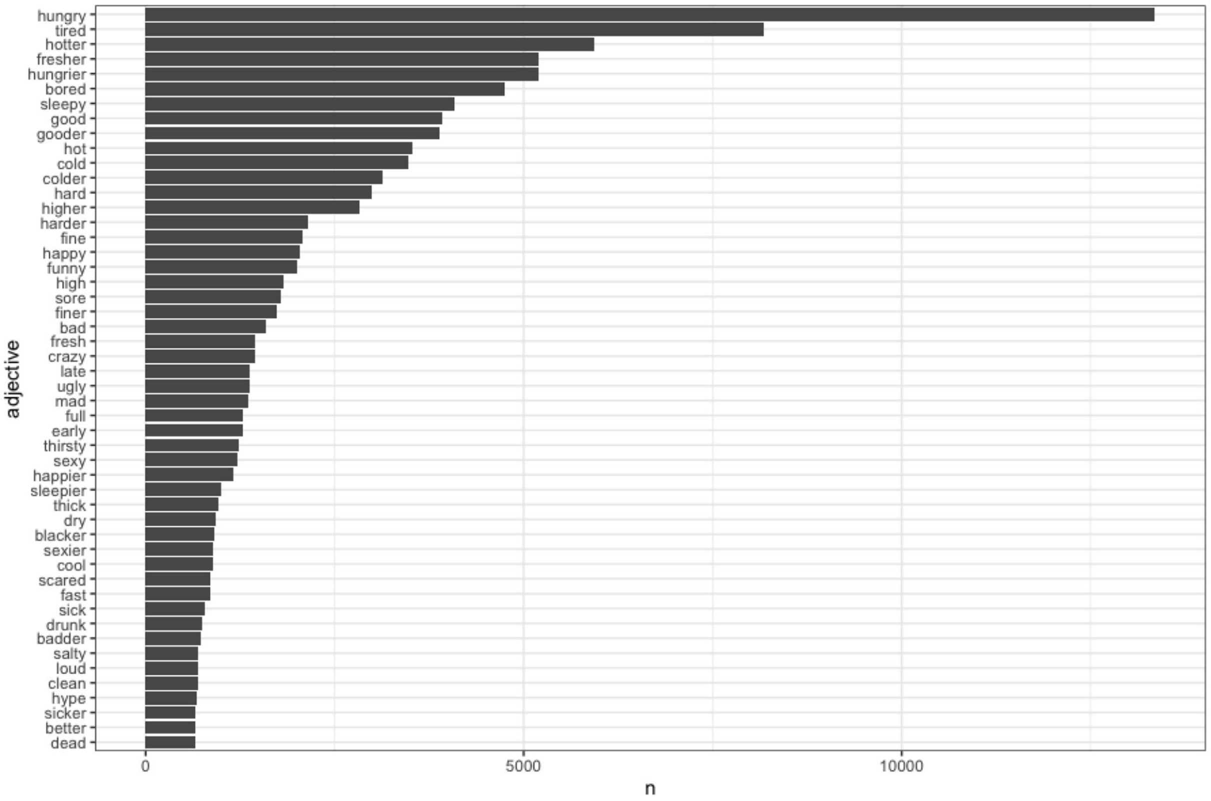


FIGURE 4
Fifty most common adjectives (normalized spelling).

TABLE 1 Results of logistic regression on the full data set.

	Estimate	SE	*z* value	Pr(>|*z*|)
(Intercept)	0.43	0.01	37.15	0.00***
thStoppingTRUE	0.19	0.02	12.27	0.00***
Vowel: e	−1.41	0.01	−105.10	0.00***
Vowel: i	−1.17	0.03	−37.19	0.00***
Fortition	−0.82	0.01	−68.59	0.00***
Complexity: reduced	−0.10	0.02	−4.99	0.00***
Lemma frequency	0.05	0.00	9.97	0.00***
Adj: good (TRUE)	1.25	0.02	58.46	0.00***

The *** symbol indicates the significant at the 0.001 level.

TABLE 2 Results of logistic regression on the subset of data comprising authors who tweeted once.

	Estimate	SE	*z* value	Pr(>|*z*|)
(Intercept)	0.86	0.02	49.64	0.00***
thStoppingTRUE	−0.02	0.02	−0.91	0.36
vowele	−1.38	0.02	−72.83	0.00***
voweli	−0.86	0.08	−11.15	0.00***
Fortitionfortition	−0.91	0.02	−49.87	0.00***
complex2reduced	−0.36	0.04	−8.94	0.00***
Scale(lemmaFreq)	0.07	0.01	9.26	0.00***
isGoodTRUE	1.00	0.04	25.54	0.00***

The *** symbol indicates the significant at the 0.001 level.

tweeted at least 10 times is presented in Table 3. The intercept is −1.52, compared to 0.86 for those who tweeted once, indicating that all things equal, those who tweeted once were likely to tweet with comparative morphology 70% of the time, whereas those who tweeted 10 or more times were only likely to tweet with comparative morphology on the adjective 18% of the time.

For this subset of the data, TH-STOPPING is once again significant, and positively associated with the presence of comparative morphology. Raising of the initial vowel was associated with loss of comparative morphology, as was fortition of the closing consonant. Orthographic complexity and lemma frequency were not significant.

TABLE 3 Results of logistic regression on the subset of data comprising authors who tweeted 10+ times.

	Effect	Group	Term	Estimate	SE	z-value	p-value
1	Fixed		(Intercept)	−1.52	0.11	−14.15	0.00***
2	Fixed		TH-stopping (TRUE)	0.54	0.08	6.51	0.00***
3	Fixed		Vowel: e	−1.64	0.08	−19.91	0.00***
4	Fixed		Vowel: i	−1.62	0.12	−13.51	0.00***
5	Fixed		Fortition	−0.46	0.09	−4.78	0.00***
6	Fixed		Complexity: reduced	−0.01	0.07	−0.22	0.83
7	Fixed		Lemma frequency	−0.02	0.02	−0.99	0.32
8	Fixed		Adj: good (TRUE)	2.21	0.06	35.23	0.00***
9	Ran_pars	Username	sd__(Intercept)	2.51			

The *** symbol indicates the significant at the 0.001 level.

Across all of the data, most common form was <than a mug>, with 82,944 tokens. Extrapolating from the coefficients in the first model, the probability of seeing comparative morphology on the adjective for this form was 0.41 or approximately two in five. The forms that are most transparently related to *than a mother*, <than a muv> and <than a muh>, have a probability of 0.61 of appearing with an adjective that exhibits comparative morphology—by far the highest probability of any forms that actually appears in the data. The least transparently related possible forms to the original comparative phrase, (e.g., <dinamub>, <dinnamud>, etc.), were predicted to have a 0.19 probability of appearing with comparative morphology. The form in the data associated with the greatest likelihood of comparative deletion is <thenna mud> which appears 60 times in these data and never with comparative morphology on the adjective. The variable associated with the largest *increase* in likelihood of comparative morphology on the preceding word was whether the preceeding word was a form of *good*, although ironically, this was not due to preference for the word *better*, but rather for *gooder*.

4. Discussion

Taken together, these results suggest that when looking at *dennamug* and not *than a mother*, we are not merely looking at creative orthography to represent spoken accent (although there is strong evidence for this in AAE as well; see Jones, 2016c, for a thorough discussion). Rather, we are also looking at evidence of lexicalization and grammatical reanalysis. Almost any nonstandard spelling of *than a m2:* is already likely to show bare morphology on the adjective, but the more additional orthographic evidence of univerbation and demorphologization—spelling the intensifier as fewer than three orthographic words, closing the syllable with an (unpronounced) <b>, <g>, or <d>, changing the initial vowel so the first syllable is no longer transparently *than*—are all associated with greater probability of encountering a bare adjective.[28] The intercept suggests that all things being equal, the probability of seeing comparative morphology with some form of *dennamug* is already only two in three, which alone is strong evidence for lexicalization and grammatical reanalysis: comparative morphology before a comparative phrase is syntactically obligatory elsewhere in AAE, and with very few exceptions, intensifiers precede the adjective they modify.[29]

In many varieties of AAE, a prenasal /æ/ can be realized as [ɛ] (Jones, 2020).[30] In the syntactically ambiguous context here, that ash-to-epsilon shift can then feed the PIN-PEN merger. Not only is this consistent with some of the written forms (e.g., <dinnamug>) and with the common pronunciations, but this is only possible if the initial syllable is no longer clearly "than" to all speakers. Similarly, the above findings are consistent with the hypothesis that *mother* is more recoverable from <muv>, (cf *muv* [mʌvə] "mother" in some varieties of AAE), and less recoverable from the non-word <mub> or from the words <mud> and <mug> which may be associated with lexical interference. Indeed, as noted above, the form most likely to exhibit comparative deletion was <thenna mud>, which is not only no longer transparently *than a mu:*, but is also spelled in such a way that each component invites lexical interference.

It should be noted that, at least on twitter, authors seem to have a high level of awareness that *dennamug* is a non-prestige form (although I'm reluctant to call it non-standard, since many

[28] Unsurprisingly, th-stopping, which is broadly productive in AAE, is not associated with loss of comparative morphology, and is not consistently significant across models.

[29] The most obvious and well-studied example is intensifier *-ass*, whose grammaticalization has been the subject of extensive study.

[30] It should be noted that for other varieties, it is indeed [æ], for instance, in New York City AAE.

appear to have strong feelings about which spelling is "correct," and there is a *de facto* emerging standard spelling). *Dennamug* has already become enregistered (Agha, 2003) for some AAE speakers (e.g., Kev On Stage fans). There is an enormous amount of linguistic awareness and playfulness in these data. While the adjectives *dennamug* appears with follow the expected Zipfian distribution, the long tail of lower frequency items has a preponderance of uncommon words from high or academic registers, often in a comparative construction that does not work in a classroom setting: *ostentatious*, *temperamental*, *rhetorical*, *vomitous*, *delectable*, *subpar*, *bowlegged*, *incognito*, *belligerent*, *inebriated*, *jovial*, *dejavu*, and *schadenfreude (dinnamug)* among many others. Moreover, authors employ comparative morphology on words that do not ever receive comparative morphology in standard English: *antisocialer*, *catholicer*, *beautifuler*, *startlinger*, *overrateder*, *sunburnter*, *negligenter*, *tirededer*, *fadeder* (i.e., "drunker"), and *confuseder*, among others. Most prominent of these is *gooder*, which occurs 3,778 times in these data, and is sufficiently enregistered that there are multiple songs with the name "gooder dennamug."[31]

There is also significant intraspeaker variation. Examining the tweets of the 32,547 individual authors who wrote at least two tweets with some form of *dennamug* in them, 16,884 used a single spelling (the most prolific, tweeting <than a mug> 244 times, but even two of the top five most prolific stuck to <dinnamug>, with 98 and 85 such tweets from the fourth and fifth most prolific, respectively. The remaining 15,663 authors who tweeted at least twice using some form of *dennamug* did made use of multiple spellings (Figure 5). There is no apparent temporal pattern, so it appears as though authors are solving the spelling problem posed by *dennamug* on the fly, and re-solving the problem each time. Three of the authors made use of up to ten distinct spellings, and crucially, they were not the same spellings across these authors (Figure 6). There is, therefore, strong evidence of both inte- and intra-speaker variation in terms of how speakers choose to represent *dennamug* orthographically, suggesting that speakers are not always certain how to phonologically or syntactically bracket the expression.

Dennamug is also exhibiting even further lexicalization and possible grammaticalization in these data. While it was not feasible to automatically parse part of speech for the full data set as POS taggers are still not at a satisfactory level, with state of the art approaches performing at ~80% on tweets (Jørgensen et al., 2016), manual inspection of the data reveal interesting avenues for future study. Not only does *dennamug* modify adjectives, but it is now able to modify adverbs:

(8) a. She is driving leisurely than a mug
b. I am employed....gainfully then a mug too

It can now modify noun phrases:

(9) a. binary thinking *(dennamug)*
b. false advertising *(dennamug)*
c. foreshadowing *(dennamug)*
d. Power trip *(dennamug)*

It can now modify prepositional phrases:

(10) My accent gonna be outchea dennamug because I'm going to be exhausted[32]

It can now modify verb phrases:

(11) a. Procrastinating dennamug
b. Back when Trick was hot I was illegally dl'ing den a mug[33]
c. That chick is line stepping dennamug
d. I'm laughing dennamug cause I'm sure they gone get rid of your favorite. Trust me
e. Projecting dennamug!!
f. Tweet watching dennamug
g. I be scanning dennamug on a day like this here.
h. Cramping dennamug

And it can even modify entire clauses and sentences:

(12) a. No weapon formed against me shall prosper. Dennamug.
b. "You one of them or one of us?" That was a loaded question dinamug. #ShotsFired

Unfortunately, there are many social factors that the present study cannot disentangle. It is clear that beyond linguistic playfulness, there is an element of Black identity construction at play for many of the authors of these tweets, and authors recruit a variety of AAE features to construct or hint at "Black" *personae* (D'Onofrio, 2020; King, 2021). Many write <fahn> for [fɑːn] "fine," representing /ay/-monophthongization, or <fye> for *fire*, capturing postvocalic /r/ deletion. Example 12a is particularly interesting because it is an ironic use of *dennamug*, which relies on the audience finding the humor in juxtaposing sacred and profane, within a Black American Christian context, for comedic effect: The sentence is a reference to the 1996 Fred Hamilton song "No Weapon," which is itself an adaptation of Isaiah 54:17, likely from the New King James Version translation

[31] E.g.: https://soundcloud.com/blackmuzik/feelin-gooder-dennamug-feat.

[32] outchea < *"out here"*. Note, also, that the author is discussing what Labov refers to as the Sociolinguistic Monitor (Labov et al., 2011).

[33] dl < *"download."*

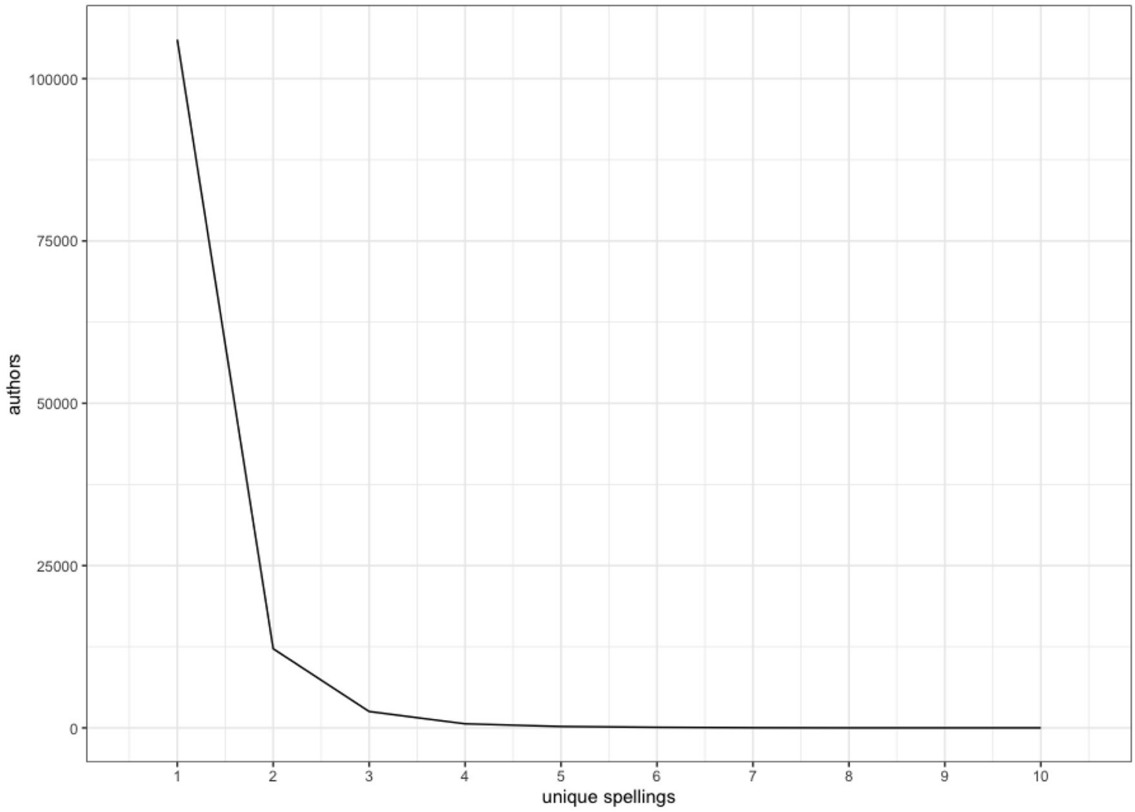


FIGURE 5
Number of authors by number of unique spellings.

"No weapon formed against you shall prosper"[34], and here *dennamug* is replacing the expected affirmation, *amen*. Other low-frequency, difficult-to-study, yet nevertheless attested AAE phenomena abound, for instance, the shift from /t/ to /k/ in initial sCr clusters (see, e.g., Bailey and Thomas, 1998, p. 89):

(13) a. I'm hongrier than a mug in this class. Lord, please give me skrenf [strength][35]
b. Hot than a mug out here in these skreets... [streets]
d. it's skrowng than a muh [strong][36]
man that game skressful thanamug [stressful]

There is much future research that could, and should, be done on this subject. One domain for future inquiry is the possibility of age grading (Hockett, 1950; Labov, 1994; Tagliamonte, 2012). One friend of the author asked "are people still saying that?" when told about this study, and indeed, there are suggestions in the data that age grading may be a very real possibility, e.g.:

(14) Just realized I'm too grown to be saying dennamug

Future work could make use of targeted elicitation, and of surveys, to better tease apart social factors related to adoption and use of this form. It is possible that autocorrect plays a role in the strong preference for <mug> and < mud> in the written data. It should be noted that those who chose to write <than a mub> must intentionally override autocorrect. There may also be lexical interference from words like *mug* and *mud*, that a future psycholinguistic study could disentangle. One important question future research should address is the question of whether bare adjectival morphology is the result of lexicalization of *than a mother* or if it preceded and fed that lexicalization. I am unaware of academic literature on bare adjectives in comparative constructions in AAE, however the phenomenon is known by AAE speaking linguists (for instance, Hiram Smith provides the example *she fine than a*

[34] Heb. *kol-kli yutzar 'alaykha lo yitzlakh* lit. "all implements created over you will not succeed".

[35] With TH-fronting.

[36] With orthographic representation of Southern AAE phonetic realization /ɔ/.

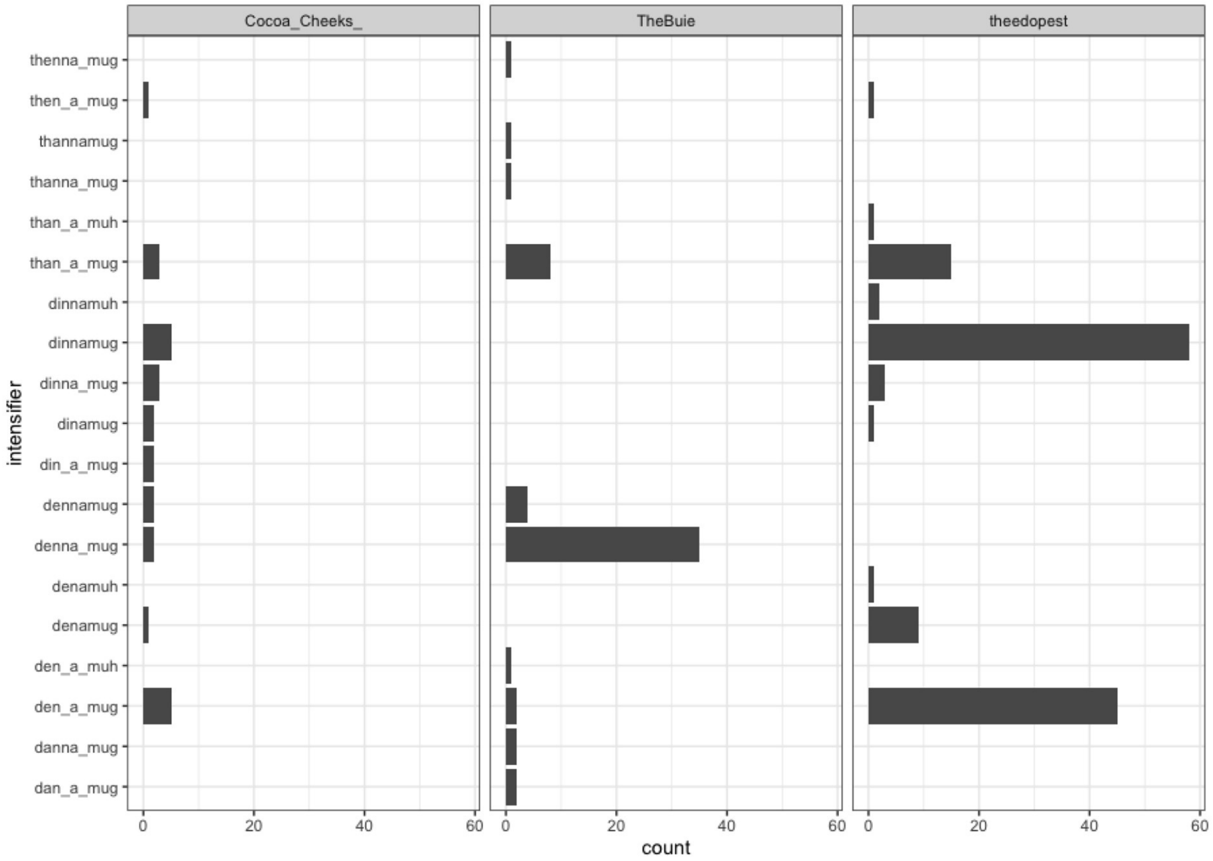


FIGURE 6
Orthographic choices for authors who used 10 different spellings.

sumbitch in a personal correspondence). The temporal ordering of these changes, whether it's bare adjectives → reduction of *than a motherfucker* to *dennamug* or vice versa, will provide important insight into pathways of grammatical change in AAE.

Nevertheless, while there is still much more to tease apart, it is clear from the above that AAE *dennamug* is an intensifier that is the result of ongoing lexicalization. Moreover, this ongoing lexicalization would have been impossible to study just a few years ago, despite being in widespread use among AAE speakers, not just because it is unlikely to appear in more formal written registers,[37] but also because the burden of proving even the existence of the phenomenon would have been to difficult for linguists using traditional methods, and the volume of data too low for analysis. The new discipline of computational sociolinguistics offers methodological innovations that allows linguists to investigate phenomena, at large scale, that we may have only heard fleetingly in the field. This broadening of methodological horizons entails a broadening of possible linguistic objects of study, and allows us to compile and study corpora of understudied languages (or linguistic phenomena) while simultaneously benefiting from linguistic transcriptions performed by the speakers themselves, rather than linguists, however well-trained. This in turn, can allow for a new window into linguistic variation and change.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/TaylorWJones/dennamug.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

[37] Or, more bluntly, in media that would not allow orthographically inventive AAE through to publication without extensive editing.

Funding

Some of this research was performed while the author was funded through the Benjamin Franklin Fellowship at the University of Pennsylvania.

Acknowledgments

TJ would like to thank Hiram Smith, Christopher Hall, Sonja Lanehart, and Brianna Meeks for helpful discussions about grammaticalization and AAE; Dennis Preston, Steven Gilbers, and Nicole Holliday for helpful feedback on an earlier form of this work presented at NWAV; Dirk Hovy, Nanna Hilton, Dong Nguyen, Christopher Purschke, and the Lorentz Center for organizing New Methods in Computational Sociolinguistics; and Jack Grieve, Rob Voigt, and Su Lin Blodgette for pushing my thinking on the intersection of NLP, social media, and sociolinguistics.

Conflict of interest

TJ was employed by CulturePoint, LLC.

The handling editor declared a past collaboration with the author.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.683104/full#supplementary-material

References

Agha, A. (2003). The social life of cultural value. *Lang. Commun.* 23, 231–273. doi: 10.1016/S0271-5309(03)00012-0

Allan, K., and Burridge, K. (2006). *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511617881

Austen, M. (2017). "Put the groceries up" comparing black and white regional variation. *Am. Speech* 92, 298–320. doi: 10.1215/00031283-4312064

Backsai-Atkari, J. (2014). *The Syntax of Comparative Constructions: Operators, Ellipsis Phenomena and Functional Left Peripheries*. Potsdam: Universitatsverlag Potsdam. doi: 10.1515/east-2014-0004

Bailey, G., and Thomas, E. (1998). "Some aspects of African-American vernacular English phonology," in *African American English: Structure, History, and Use*, eds G. Bailey, J. Baugh, J. R. Rickford, and S. Mufwene (London: Routledge), 85–109.

Bauer, L. (1994). *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Longman.

Benveniste, E. (1971). "Subjectivity in language, in *Problems in General Linguistics*, ed M. E. Meek, Vol. 1 (Gables: FL: University of Miami Press), 223–30.

Blank, A. (2008). "Pathways of lexicalization," in *Language Typology and Language Universals*, ed G. Ungeheuer (Berlin: De Gruyter Mouton), 1596–1608. doi: 10.1515/9783110194265-049

Bleaman, I. L. (2020). Implicit standardization in a minority language community: real-time syntactic change among Hasidic Yiddish writers. *Front. Artif. Intell.* 3, 33. doi: 10.3389/frai.2020.00035

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: a critical survey of "bias" in NLP. *arXiv*. doi: 10.48550/arXiv.2005.14050

Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: a case study of African-American English. *arXiv*. doi: 10.48550/arXiv.1608.08868

Bresnan, J. (1973). The syntax of the comparative clause construction in english. *Linguist. Inq.* 4, 275–343.

Brinton, L. J., and Traugott, E. C. (2005). *Lexicalization and Language Change*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511615962

Collins, C., Moody, S., and Postal, P. M. (2008). An AAE camouflage construction. *Language* 84, 29–68. doi: 10.1353/lan.2008.0059

Corver, N. F. M. (1997). Much-support as a last resort. *Linguist. Inq.* 28, 119–164.

Cukor-Avila, P. (2001). She say, she go, she be like: verbs of quotation over time in African American vernacular English. *Am. Speech* 77, 3–31. doi: 10.1215/00031283-77-1-3

Cukor-Avila, P., and Bailey, G. (2001). The effects of the race of the interviewer on sociolinguistic fieldwork. *J. Socioling.* 5, 252–270. doi: 10.1111/1467-9481.00150

Cygan, J. (1975). Synthetical comparatives in English. *Bull. Soc. Pol. Linguist.* 33, 53–57.

Davidson, L. (2006). Schwa elision in fast speech: segmental deletion or gestural overlap? *Phonetica* 63, 79–112. doi: 10.1159/000095304

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Provo, UT: BYU. Available online at https://corpus.byu.edu/coca

D'Onofrio, A. (2020). Personae in sociolinguistic variation. *Wiley Interdiscip. Rev. Cogn. Sci.* 11, e1543. doi: 10.1002/wcs.1543

Doyle, G. (2014). "Mapping dialectal variation by querying social media," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg: Association for Computational Linguistics), 98–106. doi: 10.3115/v1/E14-1011

Eisenstein, J. (2013a). "Identifying regional dialects in on-line social media," in *The Handbook of Dialectology*, eds C. Boberg, J. Nerbourne, and D. Watt (Wiley), 368–383. doi: 10.1002/9781118827628.ch21

Eisenstein, J. (2013b). "Phonological factors in social media writing," in *Proceedings of the Workshop on Language Analysis in Social Media* (Atlanta, GA: Association for Computational Linguistics), 11–19.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *J. Socioling.* 19, 161–188. doi: 10.1111/josl.12119

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS ONE* 9, e113114. doi: 10.1371/journal.pone.0113114

Enzinna, N. (2017). "How speakers select synthetic and analytic forms of English comparatives: an experimental study," in *The Proceedings of the Linguistic Society of America* (Washington, DC), 2. doi: 10.3765/plsa.v2i0.4101

Farrington, C. (2018). Incomplete neutralization in African American English: the case of final consonant voicing. *Lang. Var. Change* 30, 361–383. doi: 10.1017/S0954394518000145

Friedman, L., and Reed, A. (2020). *The State of Linguistics in Higher Education, Annual Report*. Linguistic Society of America. Available online at: https://www.linguisticsociety.org/sites/default/files/Annual%20Report%202020%20Jan2021%20-%20final_0.pdf

Gray, F. G. (Director). (1995). *Friday*. Produced by New Line Productions and Ghetto Bird Productions.

Grieser, J. (2019). Toward understanding the n-words. *Am. Speech* 94, 409–419. doi: 10.1215/00031283-7991448

Grieve, J., Nini, A., and Guo, D. (2017). Analyzing lexical emergence in modern American English online 1. *Engl. Lang. Linguist*. 21, 99–127. doi: 10.1017/S1360674316000113

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on American social media. *J. Engl. Linguist*. 46, 293–319. doi: 10.1177/0075424218793191

Han, B., and Baldwin, T. (2011). "makn sens a# twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, OR: Association for Computational Linguistics) 1, 368–378.

Hockett, C. (1950). Age-grading and linguistic continuity. *Language* 26, 449–457. doi: 10.2307/410396

Hopper, E., and Traugott, P. J. (2003). *Grammaticalization*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139165525

Hudley, A. C., Mallinson, C., and Bucholtz, M. (2020). Toward racial justice in linguistics: interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language* 96, e200–e235. doi: 10.1353/lan.2020.0074

Izvorski, R. (1995). "A dp-shell for comparatives," in *Console III Proceedings*, eds A. Bisetti, L. Bruge, J. Costa, R. Goedemans, N. Munaro, and R. van de Vijver (The Hague: Holland Academic Graphics), 99–121.

Jespersen, O. (1949). *A Modern English Grammar*. Copenhagen: Einar Munksgard.

Jones, T. (2015). Toward a description of African American vernacular English dialect regions using "black twitter". *Am. Speech* 50, 403–440. doi: 10.1215/00031283-3442117

Jones, T. (2016a). "Aae talmbout: an overlooked verb of quotation," in *University of Pennsylvania Working Papers in Linguistics* (Philadelphia, PA), 22, 91–99.

Jones, T. (2016b). "Eem neagation in AAVE: a next step in Jespersen's cycle," in *University of Pennsylvania Working Papers in Linguistics* (Philadelphia, PA), 22, 159–166.

Jones, T. (2016c). "Tweets as graffiti: what the reconstruction of Vulgar Latin can tell us about black twitter," in *English in Computer-Mediated Communication: Variation, Representation, and Change*, ed L. Squires. (Berlin; Boston, MA: de Gruyter) p. 43–68. doi: 10.1515/9783110490817-004

Jones, T. (2020). *Variation in African American English: The Great Migration and Regional Differentiation* [PhD thesis]. Philadelphia, PA: University of Pennsylvania.

Jones, T., and Hall, C. (2019). Grammatical reanalysis and the multiple n-words in African American English. *Am. Speech* 94, 478–512. doi: 10.1215/00031283-7611213

Jørgensen, A., Hovy, D., and Søgaard, A. (2016). "Learning a POS tagger for AAVE-like language," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, CA: Association for Computational Linguistics), 1115–1120. doi: 10.18653/v1/N16-1130

Kendall, T., and Farrington, C. (2020). *The Corpus of Regional African American Language*. Eugene, OR: The Online Resources for African American Language Project.

King, S. (2021). Rethinking race and place: the role of persona in sound change reversal. *J. Socioling*. 25, 159–178. doi: 10.1111/josl.12454

Labov, W. (1994). *Principles of Linguistic Change. Volume 1: Internal Factors*. New York, NY: Wiley.

Labov, W. (2018). The role of the avant garde in linguistic diffusion. *Lang. Var. Change* 30, 1–21. doi: 10.1017/S0954394518000042

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., Nagy, N., et al. (2011). Properties of the sociolinguistic monitor. *J. Socioling*. 15, 431–463. doi: 10.1111/j.1467-9841.2011.00504.x

Langacker, R. W. (1977). "Syntactic reanalysis," in *Mechanisms of Syntactic Change*, ed C. N. Li (Austin: University of Texas Press), 57, 139. doi: 10.7560/750357-005

Lechner, W. (1999). *Comparatives and DP-structure* (Doctoral dissertations). University of Massachusetts Amherst, Amherst, MA, United States. Available online at: https://scholarworks.umass.edu/dissertations/AAI9920620

Lechner, W. (2004). *Ellipsis in Comparatives*. Berlin–New York: Mouton De Gruyter. doi: 10.1515/9783110197402

Leech, G., and Culpepper, J. (1997). "The comparison of adjectives in recent British English," in *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, eds T. Nevalainen, and L. Kahlas-Tarkka (Helsinki: Société Néophilologique), 353–373.

Lindquist, H. (2000). Livelier or more lively? syntactic and contextual factors influencing the comparison of disyllabic adjectives. *Lang. Comput*. 30, 125–132. doi: 10.1163/9789004485211_012

Miller, W. J. (2017). *Grammaticalization in English: A Diachronic and Synchronic Analysis of the "ASS" Intensifier* [PhD thesis]. San Francisco State University, San Francisco, CA.

Mufwene, S. (1998). "The structure of the noun phrase in African-American vernacular English," in *African-American English: Structure, History and Use*, eds G. Bailey, J. Baugh, J. R. Rickford, and S. Mufwene (London: Routledge), 69–85.

Pryor, R. (1976). *"Mudbone goes to Hollywood" track B2 on the LP album "Bicentennial Nigger"*. Released by Warner Bros.

Rickford, J., and McNair-Knox, F. (1994). "Addressee-and topic-influenced style shift: a quantitative sociolinguistic study," in *Sociolinguistic Perspectives on Register*, eds D. Biber, and E. Finegan (Oxford University Press), 235–276.

Smith, H. (2019). Has nigga been reappropriated as a term of endearment? A qualitative and quantitative analysis. *Am. Speech* 94, 420–477. doi: 10.1215/00031283-7706537

Spears, A. K. (1998). "African-american language use: ideology and so-called obscenity," in *African-American English: Structure, History, and Use*, eds G. Bailey, J. Baugh, J. R. Rickford, and S. Mufwene (London: Routledge), 226–250.

Tagliamonte, S. (2012). *Variationist Sociolinguistics: Change, Observation, Interpretation*. New York, NY: Wiley-Blackwell.

Thomas, E., and Bailey, G. (2015). "Segmental phonology of African American English," in *The Oxford Handbook of African American Language*, ed S. Lanehart (Oxford/New York: Oxford University Press), 403–419.

Thomas, E. R. (2007). Phonological and phonetic characteristics of African American vernacular english. *Lang. Linguisti. Compass* 1, 450–475. doi: 10.1111/j.1749-818X.2007.00029.x

van Halteren, H., and Oostdijk, N. (2012). Towards Identifying Normal Forms for Various Word Form Spellings on Twitter. *Comput. Linguist*. Netherlands J. 2, 2–22. Available online at: https://clinjournal.org/clinj/article/view/12

Wilson, J., and Gordie, B. (1957). *Reet Petite (The Finest Girl You Ever Want to Meet)*. Brunswick, NJ.

Wischer, I. (2000). "Grammaticalization versus lexicalization -"methinks" there is some confusion," in *Pathways of Change: Grammaticalization in English*, eds O. Fischer, A. Rosenbach, and D. Stein (Amsterdam: John Benjamins). p. 355–370. doi: 10.1075/slcs.53.17wis

Yuan, H., Diansheng, G., Kasakoff, A., and Grieve, J. (2016). Understanding us regional linguistic variation with twitter data analysis. *Comput. Environ. Urban Syst*. 59, 244–255. doi: 10.1016/j.compenvurbsys.2015.12.003

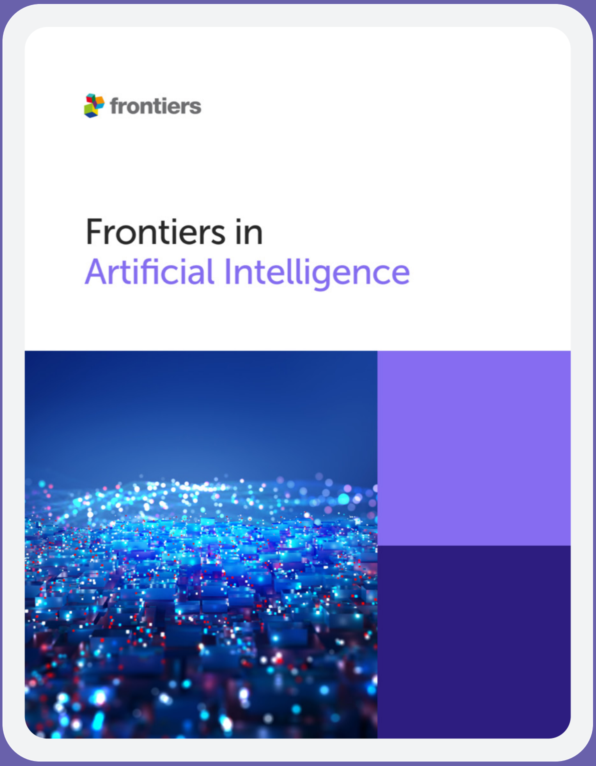
Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →


frontiers

Frontiers in
Artificial Intelligence

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact