frontiers
in Communication

ORIGINAL RESEARCH
published: 26 January 2021
doi: 10.3389/fcomm.2020.620275

Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization

Elke Teich[1]*, *Peter Fankhauser*[2], *Stefania Degaetano-Ortlieb*[1] *and Yuri Bizzoni*[1]

[1]*Saarland University, Saarbrücken, Germany,* [2]*Leibniz Institute For The German Language (IDS), Mannheim, Germany*

We present empirical evidence of the communicative utility of CONVENTIONALIZATION, i.e., convergence in linguistic usage over time, and DIVERSIFICATION, i.e., linguistic items acquiring different, more specific usages/meanings. From a diachronic perspective, conventionalization plays a crucial role in language change as a condition for innovation and grammaticalization (Bybee, 2010; Schmid, 2015) and diversification is a cornerstone in the formation of sublanguages/registers, i.e., functional linguistic varieties (Halliday, 1988; Harris, 1991). While it is widely acknowledged that change in language use is primarily socio-culturally determined pushing towards greater linguistic expressivity, we here highlight the limiting function of communicative factors on diachronic linguistic variation showing that conventionalization and diversification are associated with a reduction of linguistic variability. To be able to observe effects of linguistic variability reduction, we first need a well-defined notion of choice in context. Linguistically, this implies the paradigmatic axis of linguistic organization, i.e., the sets of linguistic options available in a given or similar syntagmatic contexts. Here, we draw on word embeddings, weakly neural distributional language models that have recently been employed to model lexical-semantic change and allow us to approximate the notion of paradigm by neighbourhood in vector space. Second, we need to capture changes in paradigmatic variability, i.e. reduction/expansion of linguistic options in a given context. As a formal index of paradigmatic variability we use entropy, which measures the contribution of linguistic units (e.g., words) in predicting linguistic choice in bits of information. Using entropy provides us with a link to a communicative interpretation, as it is a well-established measure of communicative efficiency with implications for cognitive processing (Linzen and Jaeger, 2016; Venhuizen et al., 2019); also, entropy is negatively correlated with distance in (word embedding) spaces which in turn shows cognitive reflexes in certain language processing tasks (Mitchel et al., 2008; Auguste et al., 2017). In terms of domain we focus on science, looking at the diachronic development of scientific English from the 17th century to modern time. This provides us with a fairly constrained yet dynamic domain of discourse that has witnessed a powerful systematization throughout the centuries and developed specific linguistic conventions geared towards efficient communication. Overall, our study confirms the assumed trends of conventionalization and diversification shown by diachronically decreasing entropy, interspersed with local, temporary entropy highs

OPEN ACCESS

Edited by:
Antonio Benítez-Burraco,
Sevilla University, Spain

Reviewed by:
Michael Pleyer,
Nicolaus Copernicus University in Toruń, Poland
Stefan Hartmann,
Heinrich Heine University of Düsseldorf, Germany

****Correspondence:***
Elke Teich
e.teich@mx.uni-saarland.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *22 October 2020*
Accepted: *22 December 2020*
Published: *26 January 2021*

Citation:
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. 5:620275. doi: 10.3389/fcomm.2020.620275

pointing to phases of linguistic expansion pertaining primarily to introduction of new technical terminology.

Keywords: linguistic variation, diachronic change, register, language models, entropy

1 INTRODUCTION

Language use varies according to a number of factors, from pragmatic over cognitive to social. In on-line processing, it has been shown that specific forms of variation directly serve rational communicative goals by offering ways to modulate information density in language production, and there is ample evidence that particular linguistic choices are associated with specific levels of surprisal in language comprehension (Jaeger and Levy, 2007; Levy, 2008; Schulz et al., 2016; Delogu et al., 2017; Sikos et al., 2017). It is much less clear, however, what the communicative effects might be of particular linguistic choices recurring across interactants and interaction instances.

Spontaneously occurring linguistic accommodation among interactants in on-line situations is a widely studied phenomenon—see e.g., Coles-Harris (2017); Gessinger et al. (2019); Hume and Mailhot (2013) for the phonetic level, often also referred to as convergence or alignment in interaction (see Garrod et al. (2018) for an overview) including discussion of rational communication effects (e.g., Pickering and Garrod (2004)). Here, we come from a diachronic perspective and look at possible long-term effects of interaction within a linguistic community, which we refer to as CONVENTIONALIZATION. Conventionalization is considered a prerequisite for innovation (De Smet, 2016) and a relevant component process in long-term, persistent change, as in grammaticalization (i.e., the transformation of lexical to grammatical items; Bybee (2010); Schmid (2015)).

The other major tendency to be observed in the dynamics of language use is DIVERSIFICATION. Diversification here means that a word or word form moves away from its original usage context and settles in another one. At the lexico-semantic level, this may lead to a word becoming associated with a specialized meaning (e.g., *molecule* acquiring a specialized meaning in chemistry and losing its former interchangeability with other words, e.g., *drop*). Lexico-semantic diversification typically pertains to specific socio-cultural contexts and is associated with the formation of distinctive sublanguages or registers (Ure, 1982; Halliday, 1985; Halliday and Martin, 1993; Harris, 2002). At the lexico-grammatical level, diversification means that particular words or word forms become more closely associated with specific grammatical environments, e.g., specific lexical verbs tending to be used primarily in participle form in postmodifier position (e.g., *the theory proposed by Herschel*) rather than as finite, past tense verbs. This kind of diversification may be a step towards grammaticalization, provided it spreads to other contexts and becomes more generally relevant.

We set out to show that conventionalization and diversification are reflections of one underlying mechanism: reduction of PARADIGMATIC VARIABILITY, i.e. the choices made available in a given context. To model the paradigmatic axis, we use word embeddings (Mikolov et al., 2013), weakly neural, probabilistic language models represented as vector spaces that have been used to model lexical choice in context, including lexical-semantic change (Hamilton et al., 2016; Dubossarsky et al., 2017). To capture paradigmatic variability, we calculate the entropy among words in close paradigmatic neighbourhood, based on their cosine distance in vector space. Finally, to capture diachronic variation in paradigmatic variability, we analyze change of entropy over time.

We focus here on scientific language because it is a well-studied and fairly controlled domain of discourse. Also, scientific English is a linguistically well-researched sublanguage, which allows us to link up our results with the insights of other scholars. As our data set we use a corpus composed of the publications of the Royal Society of London, spanning more than 300 years (1665–1996) Fischer et al., 2020. Nonetheless, the methodology developed here is general and can be applied to other discourse domains, registers or languages. We will show that overall, paradigmatic variability goes down over time in scientific English, indexed by entropy reduction and an overall increase of distances between words. Typically a costly process in on-line processing (Linzen and Jaeger, 2016; Lowder et al., 2018; Venhuizen et al., 2019; Tourtouri et al., 2019), entropy reduction is here shown as a diachronic process by which language use is optimized dynamically over time, keeping in check (otherwise extravagant) linguistic variation, so as to maintain communicative function.

The remainder of the paper is structured as follows. We discuss relevant related work on rational communication from an information-theoretic perspective with a view to formal, computational models of diachronic language change (**Section 2**). **Section 3** describes the overall approach, our specific methods and the data set (corpus) used. In **Section 4** we show the results of our analysis, discussing the overall diachronic trends as well as specific linguistic patterns that emerge over time showing conventionalization and diversification effects. **Section 5** concludes the paper with a summary and discussion.

2 RELATED WORK

2.1 Predictability and Uncertainty in Human Language Processing

Research on human on-line language processing in the last decade or so has shown that prediction plays a key role in human language comprehension (see Kuperberg and Jaeger (2016) for an overview). One of the crucial insights here is that SURPRISAL, the (un)predictability of an item in context, is proportional to processing effort. This is consistently supported by evidence from behavioral as well as neurophysiological studies. It has also been shown that surprisal is linked with linguistic choice, low vs. high surprisal being correlated with reduced vs. fully expanded linguistic forms (Aylett and Turk, 2004; Levy, 2008; Mahowald et al., 2013). This holds across linguistic levels, from the phonetic to the grammatical and the discourse level (Delogu et al., 2017; Lemke et al., 2017; Sikos et al., 2017; Malisz et al., 2018; Asr and Demberg, 2020).

A related notion widely applied in studies of human language processing is ENTROPY. Entropy reflects the degree of uncertainty of the outcome of an event. In this view, on-line language processing can be characterized as the incremental reduction of uncertainty about what comes next until interpretation is completed (Hale, 2001). Regarding rational communication, the question then is whether language use is adapted to minimizing the cost involved in entropy reduction and if so, what are the linguistic means available to do so. For instance, in a recent study on reading times Lowder et al. (2018) show that entropy reduction is primarily associated with increases in first fixation duration and single fixation duration, i.e., it occurs at the earlier stages of processing which are related to lexical access. As the authors explain, this gives support to the assumption that predictability effects in reading are related to some kind of preactivation of sets of probable words. But how are relevant words activated? It seems reasonable to assume that language users are aware that different contexts of interaction are associated with specific linguistic choices, e.g., formal vs. informal situations, spoken vs. written mode, field-specific domains of discourse such as sports, religion, fashion, science, etc. There is a direct link here to the notion of sublanguage or register, i.e., culturally established domains of discourse in which particular linguistic usages are more likely than others to the extent that certain options are not available at all, thus skewing the available options and reducing them altogether. This would then imply that language users, rather than operating on the full language system, have available a repertoire of linguistic subsystems tied to specific, socio-culturally established situational contexts that are activated as needed. Recent work on conversation corroborates this assumption, e.g., Hawkins et al. (2020) show that as interlocutors agree on a common ground, the set of linguistic options is effectively reduced. As specific contexts become more established socio-culturally over time, interlocutors' developing preferential choices and reducing options according to context can be considered an optimization process acting on the language system diachronically. Apart from benefits for on-line processing, entropy reduction may partly be motivated by better learnability. For instance, De Deyne et al. (2018) in a set of word neighbour generation tasks found that learners are attuned to paradigmatic relations. Or Cornish et al. (2016) in a simulation of cross-generation transmission found a cumulative increase in chunk-based structure reuse, leading to more accurate recall in learning and better memory of new structures (see also Isbilen and Christiansen (2020) for a wider overview).

2.2 Diachronic Language Change

Language use is inherently dynamic and exposed to two major pressures: innovation and conventionalization. Innovation is associated with a need for expressivity under changing socio-cultural conditions (Nettle, 1999; Labov, 1994; Labov, 2001; Trudgill, 2008), with direct reflexes in lexico-semantics. While the long-term effect on the language system (here: the lexicon) is overall expansion, repeated interaction between speakers/writers leads to convergence in language use among interactants and conventionalization sets in. For example, there may be multiple expressions denoting the same object that are used interchangeably for a while (e.g., automobile, car) until one of them dominates or even ousts the other. Or, items become conventionally associated with a particular meaning, occupying an interpersonal (e.g., adverbs expressing stance) or a textual function (e.g., adverbs functioning as discourse connectors). While convergence may also be socially determined (prestige, peer pressure) we will show that it results in a reduction of linguistic variability.

Effects of innovation and conventionalization are also encountered at the lexico-grammatical level, where items may leave their traditional contexts and acquire new (grammatical) functions or converge on one function over time. A specific example is examined in De Smet's study (2016) of the noun *key*, showing how it moved to other contexts and adopted different functions and ultimately came to be used as predicative adjective. The more general mechanism proposed by De Smet is that for innovation to occur, items need first to be conventionalized in one grammatical context, thus improving their retrievability, and subsequently become available in different, yet closely related grammatical contexts. Studies like De Smet's are set in usage-based grammar which holds that grammar is the cognitive organization of one's experience with language. Against this background, conventionalization is said to enhance retrievability (see Bybee and Hopper (2001); Bybee (2010); Schmid (2015); De Smet (2016)). In the longer term, change in language use may result in grammaticalization, i.e., particular lexicalizations become autonomous from other lexicalizations or lexical items become grammatical items (Bybee, 2010, 107). Often grammaticalization affects chunks or sequences of items (i.e., constructions), which may get reduced as their frequency of use increases. An example from the history of English is the *-ed* suffix as a reduction of the preterite *dedu* (*I did*), occurring shortly after the Germanic branch separated from the remainder of Indoeuropean (Speyer (2007)). Another example from more recent times is *gonna* from *going to*, a future marker that developed from the lexical verb *go* (Leech et al., 2009; Mair, 2017). Once chunks are reduced, they become easier to use in new contexts, thus concluding the cycle of innovation and conventionalization. Importantly, this cycle is a self-feeding process fired by frequency of use at various stages (cf. (Bybee, 2010, 109). Grammaticalization is thus not the end-point of a change but importantly, it opens up new possibilities for interpretation by pragmatic inference, e.g., in the case *going to/gonna* the habitual inference of 'intention' (cf. also Lehmann (1995); Newmeyer (2001); Traugott and Dasher (2002); Eckart (2012)).

We will show instances of this cycle in our own data in **Section 4** below, including chunks/constructions deriving from verbs. For instance, there are some polyfunctional verb forms that shift between lexical and grammatical uses, e.g., the past participle form of *provide* or the present participle form of *consider*. Both come to be used as conjunctions (*provided that*, *considering that*), including a reduced form without *that*. Diachronically, the grammatical use of these forms in our data set becomes dominant over the lexical one and they rise in frequency. With the reduced version, there is again a rise in frequency of occurrence and a strong syntagmatic fixation, e.g., in the case of *considering* on a definite noun phrase. Interestingly, there is no inverse process of 'lexicalization' (grammatical items to lexical items), and grammaticalization is irreversible (for a discussion see Haspelmath (1999)), which is consistent with the view that grammar (structure, constraints on linearization) enables code optimization.

While many interesting and relevant insights come from the recent works on the underlying mechanisms, conditions and possible reasons of linguistic change, there are also some limitations. First, predominantly frequency-based approaches may risk to rely too much on the sometimes fairly weak link between (change in) frequency and cognitive processes (for a discussion see Arppe et al. (2010)). According to the more recent information-theoretically based rational accounts of human language processing it is not so much frequency directly that indexes processing effort but information content (measured e.g. by surprisal). The perspective of information, while potentially very fruitful, has so far only rarely been adopted in language change. For example, in a study of the conditions of sound change Hume and Mailhot (2013) show that phonologization tends to affect elements linked to extreme degrees of surprisal and that both very low or very high surprisal exhibit low contributions to predicting outcomes in a system, i.e., to entropy reduction. In our own work, we have forwarded the hypothesis that scientific English has diachronically evolved towards an optimal code for communication among experts (Degaetano-Ortlieb and Teich, 2019; Bizzoni et al., 2020). Using information-theoretic measures (relative entropy, average surprisal), we have found that scientific English drifts away from general language over time, indicated by relative entropy (Kullback-Leibler Divergence) due to distinctive syntactic usage at clause level and a preference for complex nominal expressions. Degaetano-Ortlieb and Piper (2019) confirm this trend for the humanistic domain of literary studies using the same methodology. These studies provide support to former descriptive as well as corpus-based works such as Halliday and Martin (1993) or Biber and Gray (2016) and add the specific aspect of communicative concerns in diachronic language change. Second, existing works often focus on specific constructions or items that are hand-selected (e.g., on the basis of frequency-based corpus analysis). While compelling for individual linguistic phenomena, a wider perspective on change in the language system is prevented with a phenomenon-driven approach and generalizations are thus impeded. To be able to adopt a combined systemise+ perspective, a more exploratory, data-driven approach that can be naturally adapted to diachronic analysis is called for.

2.3 Computational Models of Language Change

The most common approach to modeling diachronic change are distributional models and more specifically word embeddings, which rely on the fact that words with related meanings occur in similar contexts (cf. Lenci (2008)). Technically, computed on the basis of a corpus, a co-occurrence matrix of words is built up from which a vector space is generated. Once such a space is generated it is possible to compute the distributional difference between two words as their distance from each other. A common measure to quantify distance is computing the cosine of the angle between two words. In a diachronic scenario, changes in cosine distance between words in a vector space indicate that words shift in use. Gulordava and Baroni (2011) were among the first to show large-scale lexical-semantic change based on the Google NGram corpus using this method.

Naturally, the method of defining and analyzing the topology of words in a vector space determines which kinds of distributional behaviours we are able to observe. For example, Hamilton et al. (2016) show that focusing on changes in a word's close neighbourhood highlights cultural shifts in word meaning while focusing on its global change with respect to the overall topology of the space highlights linguistic shifts in word usage. Similarly, Dubossarsky et al. (2016) show that the grammatical categories words belong to play an important role in the way they shift through diachronic spaces. In our own work, we have observed that topological shifts in diachronic word embeddings are effects of the tension between lexical and grammatical changes (Bizzoni et al. 2019; Bizzoni et al., 2020). Here, we build on these insights and specifically inspect tendencies towards grammaticalization. Closely related to the approach we pursue here in that distributional models are employed to model the dynamics of language use with a focus on grammar rather than lexis are recent works by Gries and Hilpert (2008); Hilpert and Perek (2015); Perek (2016). For a more comprehensive overview on the use of word embeddings for diachronic study see also Kutuzov et al. (2018).

2.4 Cognitive Relevance of Word Embeddings

From a processing perspective, some recent work highlights correlations between distributional properties of words and cognitive indices: distributional semantic models seem to mirror some aspects of cognitive lexical organization. Specifically, Abnar et al. (2018) explore how helpful different types of word representation are to a machine learning system for predicting the brain patterns activated by concrete nouns (as reported by fMRI), and find that neural word embeddings are better than count-based and association-based word models in predicting which brain voxels specific nouns will activate. Schwartz and Mitchell (2019) find that neural word embeddings can be predictive of language-elicited encephalography (voltage fluctuations through the scalp, another proxy for brain areas activation) in the sense that they can be used as input for a machine learning system that tries to

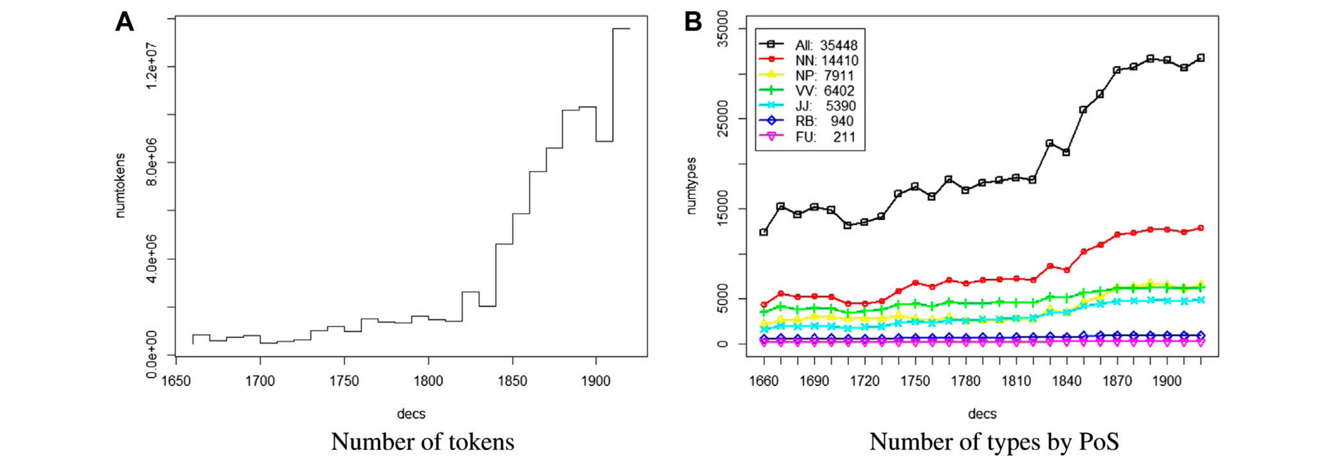

FIGURE 1 | (A) Number of tokens and **(B)** number of types by PoS per decade.

predict which scalp sensors will be most activated by given words. In Hollenstein et al. (2019) word-level cognitive data from different modalities, including eye tracking, EEG and fMRI, were converted into vectors that were fit to different types of word embeddings by neural regression with one hidden layer and linear activation. The authors found overall strong correlations between distributional and cognitive representations. Distance between words in vector space, as measured by cosine distance, also appears to weakly, but positively, correlate with human reaction times in lexical decisions and naming tasks (Auguste et al. (2017)).

3 DATA AND METHODS

3.1 Data

The data set we use is the Royal Society Corpus (RSC) v6.0, covering ca. 250 years of scientific articles (1665–1929), roughly spanning the late Modern period (ca. 1700–1900). This period is linguistically interesting insofar as many new registers emerge, including the scientific one, due to increasing societal diversification. The corpus comprises 91.2 million tokens over about 462.000 types and has been split into 27 decades, with the number of tokens per decade ranging between 455.351 and 13.583.475. The corpus is tokenized, lemmatized and tagged with parts of speech. The larger part (noncopyrighted material) is available under a Creative Commons license and accessible via a web concordance. For a comprehensive description of the RSC see Fischer et al. (2020).

An important characteristic of the RSC is the imbalance in size across time periods, the more recent periods being much larger than the earlier ones (**Figure 1A**). Naturally, the increase in number of tokens is reflected as an increase of the number of types overall (considering only types that occur at least 50 times in the corpus), shown in **Figure 1B** by part-of-speech (NN = noun, NP = proper noun, VV = lexical verb, JJ = adjective, RB = adverb, FU = function word). 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 RSC is the most comprehensive and largest diachronic corpus of English Scientific writing to date. It is a valuable resource not only for linguistic analysis but also for cultural studies, since it reflects different stages of professionalization in scientific writing and publication. For example, previous studies using the corpus have shown that there is a clear push around 1750 from conceptually oral to written production (Degaetano-Ortlieb and Teich, 2019). The early documents are letters to the editor characterized by a reporting style and only towards the end of the 18th century the research article develops to be the standard form of written knowledge transmission. The RSC comes with rich meta-data, including time period, authors and topics, thus offering interesting variables of analysis to linguists as well as historians.

3.2 Computational Modeling

The word embedding model we use are structured skipgrams (Ling et al., 2015), an extension of skipgram word embeddings introduced in Mikolov et al. (2013). Whereas skipgrams represent the left/right usage context of a word as a bag of words, structured skipgrams represent each position in the context separately. For characterizing content words skipgrams and structured skipgrams seem to fare equally well, but structured skipgrams do better for characterizing function words. This is crucial in the present context because we want to trace shifts in word usage from lexis to grammar.

For computing period-specific word embeddings that are aligned with each other, we have experimented with two variants of the approaches presented by Dubossarsky et al. (2017) and Fankhauser and Kupietz (2017). Training for the first period is either initialized randomly (Option 1), or on "atemporal" embeddings trained on the complete corpus (Option 2). All subsequent periods are then initialized with the embeddings of their previous period. For the random initialization option, embeddings for the complete corpus are initialized with embeddings for the last period.

For words with enough support these two options seem fairly equivalent. However, low frequency words can behave rather differently: with random initialization low frequency words tend to be rather arbitrarily concentrated in the center of the space for the first few periods. Corpus initialization avoids this, but then the positioning of low frequency words may not really reflect their actual usage during the first few periods. Likewise, random initialization may bias the representation of low frequency words for the complete corpus by the representation of the last period. Moreover, random initialization also leads to partially erratic movement in the space over time, evident by a larger average distance of word embeddings over time. Thus for the actual analysis in this paper, we stick to Option 2. As an extra measure we filter out low frequency words.

Initializing on larger corpora and fine-tuning on the datasets of interest is a widespread technique to counter data scarcity in both classic (Xu et al., 2015; Rothe et al., 2016; Kim et al., 2020) and contextualized word embeddings (Li and Eisner, 2019), especially for so-called down-stream tasks (Babanejad et al., 2020), i.e., applications to evaluate a model such as automatic classification, paraphrase detection or information retrieval. A similar approach was also recently used to stabilize word embeddings trained on diachronic (albeit contemporary) data (Di Carlo et al., 2019).

3.3 Measuring Diachronic Change

Our focus is on diachronic shifts in paradigmatic variability, i.e., the degree of choice in a given context/set of similar contexts, where sinking paradigmatic variability is an index of increasing conventionalization and possibly grammaticalization. Based on word embeddings, a simple measure for the paradigmatic variability of a word is the number of its close neighbours within a given radius. We employ a more refined measure that weights words x_i in the neighbourhood C_x of a word x by their frequency $\mathrm{freq}(x_i)$ and by their cosine similarity $\cos(x_i, x)$ to x[1]. On this basis, we can estimate the probability $\mathrm{p}(x_i|C_x)$ that a word x_i is chosen instead of word x. More frequent and closer words x_i get a higher probability. The paradigmatic variability is then defined as the entropy over this probability distribution:

$$\mathrm{pvar}(x) = \mathrm{H}(\mathrm{P}(.|C_x)) = -\sum_{\cos(x_i,x)>\theta} \mathrm{p}(x_i|C_x)\log\big(\mathrm{p}(x_i|C_x)\big)$$

$$\text{with } \mathrm{p}(x_i|C_x) = \frac{\cos(x_i,x)\mathrm{freq}(x_i)}{\sum_{x_j}\cos\big(x_j,x\big)\mathrm{freq}\big(x_j\big)}$$

A word with many close, rather uniformly distributed neighbours thus has high paradigmatic variability. For the threshold θ we have experimented with values between 0.7 and 0.6, settling on 0.6, which–based on inspection–gives sensible neighbourhoods overall. Moreover, we only consider a maximum of 30 neighbours.

TABLE 1 | Correlations between measures.

	mdist	nn	pvar07	pvar06
mdist	1.00	−0.76	−0.82	−0.70
nn	−0.67	1.00	0.53	0.63
pvar07	−0.84	0.47	1.00	0.61
pvar06	−0.65	0.70	0.56	1.00

Table 1 shows the correlations between the mean distance between a word and its 30 nearest neighbours (mdist), the number of neighbours (nn) with cosine similarity greater than 0.6, and the paradigmatic variability with $\theta = 0.7$ (pvar07) and $\theta = 0.6$ (pvar06). The upper diagonals give the Pearson correlation, the lower ones Spearman rank correlation. All correlations are calculated for each decade individually and then averaged. As we can see mean distance is strongly negatively correlated with all measures of paradigmatic variability.

Distance, paradigmatic variability and frequency can then be used to explore the diachronic word embedding space. For example, we may inspect specific pairs or sets of words that exhibit significant increases in topological distance, thus indicating lexico-semantic diversification and specialization in meaning, one of the reasons for reduction of paradigmatic variability. For some examples see **Table 2**. For instance, *drop* and *molecule* or *part* and *particle* are fairly close in topological space in earlier centuries and move apart in later centuries, clearly separating the more general from the more specific meaning. Similarly, we can find candidates for shifts from lexical usage to grammar, such as *owing to*.

In the following section we analyze the diachronic word embedding space in more detail, both in terms of general diachronic trends (**Section 4.1**) and in terms of the contributions to the general trends by specific word classes (**Section 4.2**), specifically focusing on paradigmatic variability and its link to communicative efficiency.

4 ANALYSES

4.1 Macroanalysis: Overall Diachronic Trends

The overarching diachronic trend consists in the expansion of the word embedding space manifested in an overall increase in the distances between words. This trend is continuous and independent of token frequency or whether a word is used continuously over time or not. **Figure 2** graphically displays the diachronic development, distinguishing between lexical words (upper points) and function words (lower points). As can be seen, the overall trend of increasing distance involves predominantly the lexical words while the function words stay diachronically stable. This is what would be expected: grammatical change is slow and function words are fairly inert, while lexis is very agile and changes in lexical usage occur at a fast rate.

The overall increase of distances between words is a reflection of the increase in types over time in the Royal Society Corpus (see

[1] For the word x, $\cos(x,x) = 1$. We have also experimented with mapping the cosine similarity to a Gaussian distribution with a standard deviation estimated from the overall distribution of distances. This gives similar overall results, but tends to be too permissive in including spurious neighbours.

TABLE 2 | Examples of word pairs with changing cosine similarity over time. We present three cases of increasing distance (indicating diversification) and one case of decreasing distance (indicating conventionalization).

Word pair	Word 1	Word 2
drop–molecule Early distance: 0.41	. . . child hath the small pox, the child is found to have them too: Though not one **drop** of the mothers blood passes into the child that the membranes and . . . (1670s)	. . . the vessels appears to have such a quantity of air intimately mixed with every **molecule**, globule, or particle of it, the whole compound according to the . . . (1730s)
drop–molecule Late distance: 0.68	. . . pressure the potential required to cause a discharge from the surface of a **drop** of water at the end of a capillary tube exceeds, though only by a few . . . (1920s)	. . . differential equation of motion is developed for the rotations of a **molecule** with two degrees of freedom, a permanent magnetic moment and a moment. . . (1920s)
part–particle Early distance: 0.42	. . . to labour after a way, whereby the parts of glass may be comminuted into such small **parts**, as to touch one another in many points, and that then malleable . . . (1660s)	. . .is means, and the earth shows quite a new thing to us, so that in every little **particle** of its matter, we may now behold almost as great a variety of creatures . . . (1660s)
part–particle Late distance: 0.77	. . . in 100,000 and, since the metal is in contact with the marble over only a small **part** of its surface, the probable error due to the base cannot exceed . . . (1920s)	. . . to the channel, the distance from the side, the longitudinal velocity of a **particle** there, and the height of the free surface above its . . . (1920s)
success–happiness Early distance: 0.47	. . . He particularly describes those, which he chiefly made use of with good **success**, from the prescriptions of the college, and of Sr. Theod. Mayern. . . (1670s)	. . . done that, he proceeds to consider the advantage of this doctrine, and its **happiness** in explicating many phenomenon, hardly explicable without it; . . . (1670s)
success–happiness Late distance: 0.58	. . . which is unique in our method, were to fail, the method would also fail: Its **success**, now to be shown, implicitly carries with it the uniqueness of the . . . (1920s)	. . . prefixed to his little book on diamonds was an indication of the domestic **happiness** which throughout accompanied his long and active career . . . (1920s)
due–owing Early distance: 0.35	. . .Life, he is of opinion, that this niter, mixed with the sulfurous parts of the blood, causes a **due** fermentation, which he will have raised, not only in the heart alone, but immediately in the. . . (1660s)	. . . hath made no thorough investigation of any plant, and left a very great number of them untouch't, **owing** also much of what he knew to the egyptians that euclid lived a while in aegypt, a country . . . (1670s)
due–owing Late distance: 0.26	. . . thus deducible at once from the integral equation, is especially useful in giving the distant field - **due** to the two discs the total charge on each disc is evident, and the exact value is given later . . . (1920s)	. . . obtained by using a control frequency of 2,000 cycles per second, but this idea was not pursued **owing** to difficulties in constructing a highly accurate and permanent maintained tuning fork or . . . (1920s)

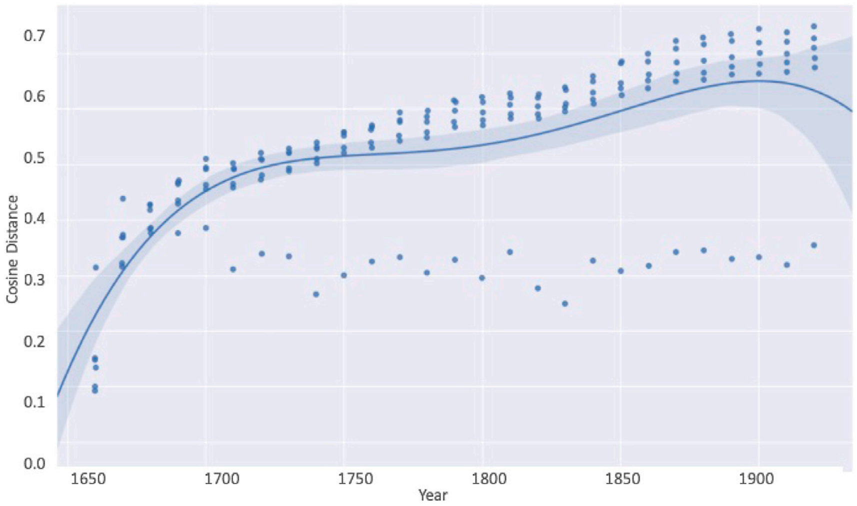

FIGURE 2 | Average cosine distances of words between randomly chosen groups of words for lexical words **(upper points)** and function words **(lower points)**.

again **Figure 1** above), on the one hand, and, as we will show in **Section 4.2**, of diversification in word usage. Again, function words (FU: determiners, prepositions, conjunctions, pronouns, and auxiliary/modal verbs) are the most stable, the number of types hardly changes over time. The increase affects mostly the lexical words and is distributed unevenly across parts of speech with nouns (NN) showing the largest increase. This indicates that unsurprisingly nouns are the primary hosts for lexical innovation and vocabulary expansion in this domain like in other domains. Note that despite the increase in types, the overall unigram entropy as well as the entropy per major part-of-speech remain remarkably stable, as shown in **Figure 3B**. We take this as a first indication that some mechanism for maintaining communicative function must be in place.

Correlating with overall increasing distance, paradigmatic variability decreases over time as a general trend. **Figure 4** shows mean distance and paradigmatic variability by major parts of speech. Function words (FU) and adverbs (RB) are

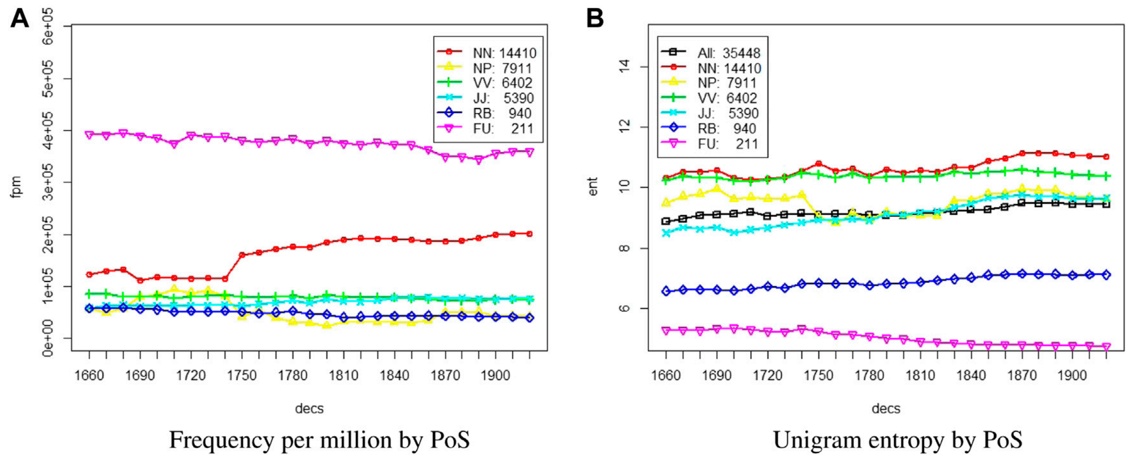

FIGURE 3 | (A) Frequency per million and **(B)** unigram entropy by part-of-speech and decade.

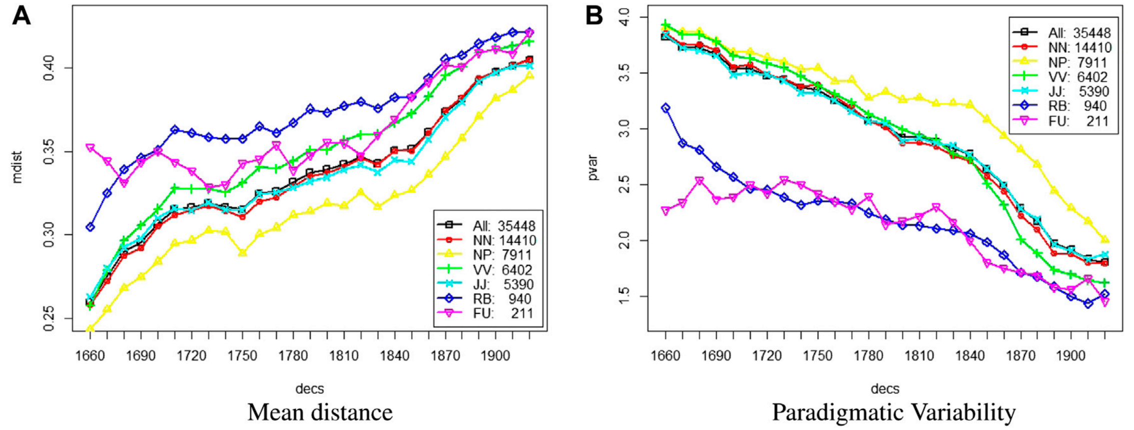

FIGURE 4 | Diachronic trends by major parts of speech in terms of **(A)** Mean Distance (mdist) and **(B)** Paradigmatic Variability (pvar) by decade.

more distant to their neighbours and have lower overall paradigmatic variability. Proper nouns (NP) in general have high paradigmatic variability.[2] Nouns (NN) and adjectives (JJ) have rather similar paradigmatic variability. Finally, verbs (VV) start out with a slightly higher paradigmatic variability than nouns, but end up with lower variability almost at the level of adverbs and function words.

Figure 5 compares the diachronic development for nouns (NN) and different verb forms. Verbs are generally more distant from their neighbours than nouns, but in terms of paradigmatic variability they are less clearly separated. While verbs start out at the same level or even at a higher level of variability, participles (VVG and VVN) and verbs in past tense (VVD) end up at lower variability, whereas verbs in base form or present tense (VV) have about the same variability as nouns. This is again an indication of diversification in usage, possibly showing a separation of grammatical and lexical uses of certain verb forms, some of them conventionalizing and moving to the grammatical end (VVG, VVN, VVD) and others staying at the lexical end (VV).

What is also noteworthy here is that verbs in base form and present tense (VV) as well as nouns are in the high frequency range, while the participle and past tense forms are in the mid-to-lower frequency range. As mentioned above, frequency plays an important role in conventionalization and grammaticalization. As we will show in **Section 4.2** below, it is the mid-to-lower frequency items that are susceptible to change by conventionalization/grammaticalization while the high-frequency ones (such as function words) are fairly immune to change.

To analyze these macroanalytic trends further, we need to inspect in more detail the different linguistic patterns that lie behind paradigmatic variability reduction, again considering the interplay with frequency and distance.

[2] This result is intuitive because names are high entropy items.

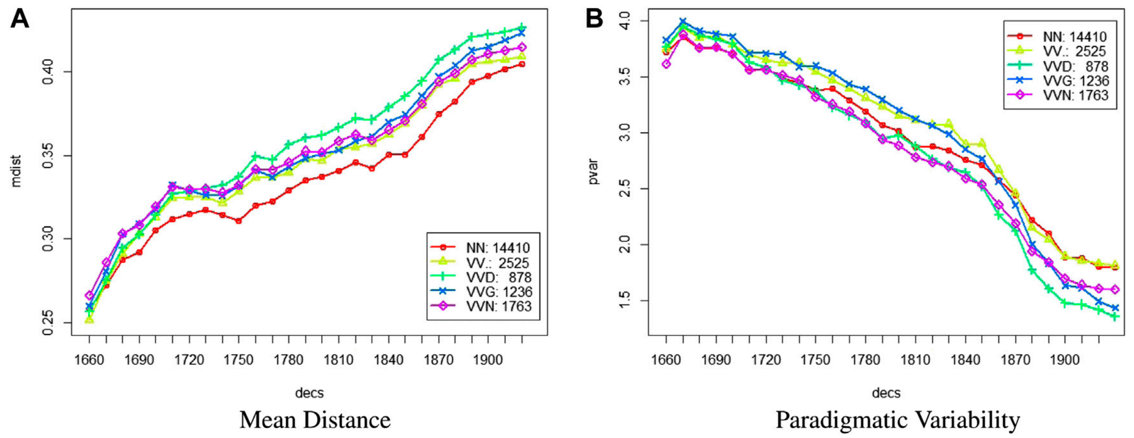

FIGURE 5 | Diachronic trends of nouns (NN) vs. verbal parts of speech (VVG, VVD, VVN) in terms of **(A)** Mean Distance (mdist) and **(B)** Paradigmatic Variability (pvar) by decade.

4.2 Microanalysis: Linguistic Patterns of Paradigmatic Reduction

We observe two main (non-exclusive) mechanisms for limiting paradigmatic variability over time: CONVENTIONALIZATION—a word becoming the dominant choice within its neighbourhood by frequency (convergence), possibly replacing other, alternative words (substitution)—and DIVERSIFICATION, i.e., words within a neighbourhood becoming more distant, possibly leading to a split into two or more neighbourhoods.

As indicated by the overall trends, different parts of speech have different roles in diachronic change, siding either with the lexico-semantic or the lexico-grammatical aspect of change. It is therefore instructive to look at the lexico-semantic and the lexical-grammatical contributions to diachronic shifts in paradigmatic variability individually.

4.2.1 Paradigmatic Reduction Pertaining to Lexico-semantic Items

As an example of conventionalization by substitution, **Figure 6** shows the frequency development of *oxygen* in comparison to its closest neighbours[3]. The former term *phlogiston*, denoting the hypothetical substance released during combustion, is substituted by (French) *oxygene* as the actual substance added during combustion, co-existing for a while with the variant *oxygen* which finally takes over. This second kind of substitution also occurs for other names of chemical elements which are close neighbours of *oxygene*, e.g., *hydrogen* and *nitrogen*. As a result of this conventionalization, *oxygen* also becomes very productive in word formation to denote processes (*oxidize*), properties (*oxidative*), molecules (*oxyhydrogen*), etc. Altogether there exist almost 50 different words derived from *oxy* in the RSC. This is a prime example of conventionalization enabling innovative linguistic uses.

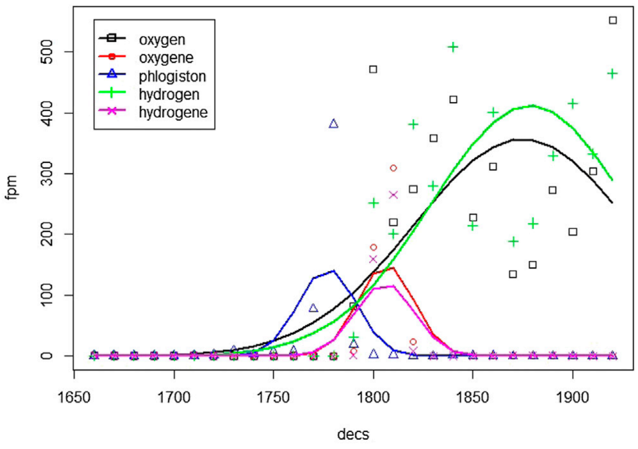

FIGURE 6 | Conventionalization by substitution for *oxygene*

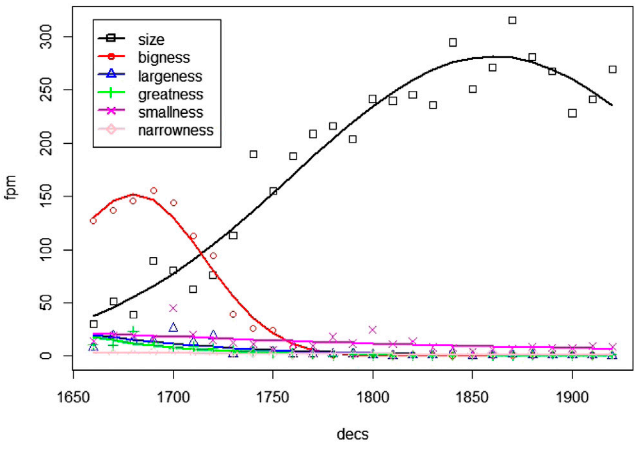

FIGURE 7 | Conventionalization by convergence on *size*.

[3] In **Figure 6** through **9** the diachronic change in relative frequency is fit with a generalized linear model with a binomial link function, whereas change in paradigmatic variability is fit with a linear model, both of degree 2.

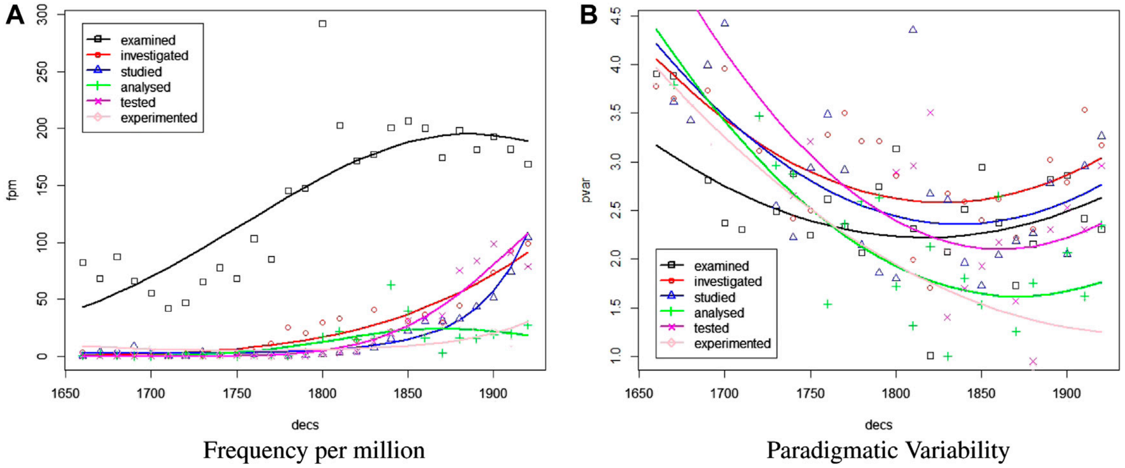

FIGURE 8 | Diversification of examined. **(A)** Frequency per Million and **(B)** Paradigmatic Variability.

More generally, when a word becomes more dominant it can substitute a whole group of words. **Figure 7** shows the frequency development of close neighbours of *size*. Initially, the dominant choice is *bigness*, but there exist a number of other choices to express various aspects of size. Then, around 1750, the general term *size* becomes the dominant choice, and all other choices become fairly rare. Similarly to *oxygen* above, *size* becomes productive in word formation, in particular as adjective (*sized*, *medium-sized*, *full-sized*). This is another example of conventionalization enabling innovation.

As an example of diversification, see again the use of *drop* and *molecule* (**Table 2**), which start as close neighbours and become clearly separated from each other when *molecule* acquires a specific meaning and becomes a close neighbour of *atom*. Here again, *molecule* becomes productive in word formation, especially as adjective (*molecular*, *bimolecular*, *intermolecular*). Similarly, *small part* and *particle* appear in the late 17th century to be virtually interchangeable, but become quite distant as *particle* starts to represent subatomic particles only (this process is already visible by the 1920s).

As an example of diversification pertaining to verbs, **Figure 8** plots the five closest neighbours of *examined*. As shown, while *examined* starts out as and remains the dominant choice by frequency, it does not substitute other choices. On the contrary, *investigated*, *studied*, and *tested* become relatively frequent choices after about 1800, while the frequency of *examined* levels out. Thus, until 1800 the distribution of their frequencies becomes less uniform leading to a decrease of paradigmatic variability. But after 1800 the frequency distribution becomes more uniform and accordingly the paradigmatic variability increases.

4.2.2 Paradigmatic Reduction Pertaining to Lexico-grammatical Items

Especially interesting from the point of view of communicative utility are trends affecting the grammatical side of words, possible leading to grammaticalization. Grammar being the most efficient linguistic encoding, any move in this direction is beneficial from the point of view of communication. Given the known paths of grammaticalization, what we look for here are words or word forms that adopt another function and split away from their dominant lexical neighbourhood moving to a grammatical neighbourhood. To find candidates involved in such shifts, we inspect words by their paradigmatic variability score, where lower entropy and greater mean distance over time are again indicators of diversification. As we will see, items may not go the full way from lexical to grammatical or they may form a new category. If an item grammaticalizes, it may be used more frequently and productively (similar to the behavior of lexical words participating in derivational processes as shown for *oxygen* in **Section 4.2.1**).

As shown in **Figure 5B** above, the largest contribution to decreasing paradigmatic variability comes from verbs in present participle form (VVG) (mean pvar: −1.34), past participle (VVN) (mean pvar: −1.80) and past tense (VVD)) (mean pvar: −1.24). To show the diachronic mechanism at work, we inspect 15 VVGS with fpm > 30 (from altogether 115 types with fpm > 30): the five with the greatest decrease in paradigmatic variability, the top five with increasing pvar and five with stable pvar (< 0.9). See **Table 3** for the items selected by this procedure. Note that for pvar- (left column) we choose VVGS with rising frequency as rising frequency items are more plausible candidates for grammaticalization. The middle column pvar + contains the top five items with increasing paradigmatic variability—these VVGS are expected to remain in their lexical neighbourhoods. The right column pvars shows items with stable paradigmatic variability. Being function words (prepositions/conjunctions), they are themselves the result of a grammaticalization process, and should also stay in their (grammatical) neighbourhoods.

TABLE 3 | Paradigmatic Variability of VVGs. Top 5 with pvar- (left); top 5 with pvar+ (middle); selected 5 with pvars (right).

pvar-	pvar+	pvars (< 0.9)
Assuming	Adding	According (to)
Leading	making	Regarding
measuring	Taking	Including
Involving	Giving	Concerning
Owing (to)	Obtaining	Considering

TABLE 4 | Three closest neighbours of VVGs with decreasing Paradigmatic Variability (pvar-) by 50-year period.

Word	1675	1725	1775	1825	1875	1925
Assuming	Attributing	Adopting	Assume	Supposing	Supposing	Supposing
	Assigning	Stating	Disregarding	Assume	Assume	Assume
	Adopting	Selecting	Equalizing	Taking	Adopting	Suppose
Leading	Leads	Prolongation	Leads	Communicating	Leads	Connecting
	Unclosed	Ramifying	Led	Inosculating	Connecting	Connected
	Outlet	Wandering	migrating	Extending	Led	Leads
measuring	Estimating	Determining	Determining	Determining	Determining	Estimating
	Predicting	Estimating	Registering	Ascertaining	measure	Determining
	Determining	Sounding	Ascertaining	Calculating	Registering	Observing
Involving	Involve	Involves	Involve	Non-linear	Involve	Involve
	Involves	Involve	Involve	Transforming	Involves	Involves
	Predicts	multinomial	Unaccented	Factorials	Canceling	Requiring
Owing	Attributable	Attributable	Attributable	Attributable	Due	Due
	Ascribable	Attributed	Occasioned	Due	Consequence	Spite
	Ascribed	Imputed	Imputed	Occasioned	Spite	Attributable

Again, the pvar-VVGs (left column) are the items of interest here since they are candidates for conventionalization/grammaticalization, i.e., they should become dominant choices in a given neighbourhood or shift to another (grammatical) neighbourhood or possibly form their own neighbourhood. If they (or some of them) shift to the grammatical end, their paradigmatic variability will become stable and they become similar to the pvars items (right column).

The micro-analysis of the neighbourhood shifts for the 15 VVGs is presented in **Tables 4–6** showing their three closest neighbours per 50-year period. What can be seen is that all 15 VVGs are polyfunctional (i.e., they have lexical and grammatical items as neighbours) but pvar-, pvar+ and pvars clearly exhibit different neighbourhood patterns.

Comparing pvar- and pvar + items, we can see that among the closest neighbours of pvar + items are other word forms of the same root, e.g., *giving*: *give gives*. The closest neighbours of pvar-items instead are other *ing*-forms and for some, their neighbourhood gets clearly more confined and stable over time. For example, the neighbourhood of *assuming* has 30 close neighbours (including *supposing*, *assume*, *considering*) in the first decade, but only 13 close neighbours in the last decade, with *assuming* and *assume* dominating by frequency.

Comparing pvar- and pvars items, we see that pvars items side with *ing*-forms similar in meaning that can also be used as prepositions, e.g., the diachronically consistent neighbours of *concerning* are *regarding* and *respecting*. The clearest diachronic trend among the pvar-items is shown by *owing (to)*. *Owing to* is actually established as a preposition by the mid 18th century (or earlier) and listed in the OED under the entry of *owing*.[4] Its usage in the RSC shows that it moved closer to be a preposition in the time span considered as seen by its neighbours: diachronically, *owing (to)* lands with *due (to)*, (*as a*) *consequence* and (*in*) *spite (of)* (cf. **Table 2** above showing the decreasing distance between *owing (to)* and *due (to)*).

For *assuming* we can observe that use at sentence beginning significantly increases over time (1810: 2.76 fpm, 1900: 33.21 fpm), obviously offering a shorter alternative to finite conditional clauses (*When/If we assume x* …). See two examples of typical usage at sentence beginning, one with *assuming* plus that-clause and one with a nonfinite clauses in 1 and 2.

1) *Assuming that the distance of the source of light from the thermopile is fixed […] still, if the india-rubber rings should become a little stretched in time, or any similar accident happen, the sensitiveness of the galvanometer would vary* (On chemical dynamics and statics under the influence of light, by Meyer Wilderman, 1902)
2) *Assuming the formula given for V to hold for this value of l/B, we see that this greatest slope is […] 810* (On an approximate solution for the bending of a beam of rectangular cross-section under any system of load, with special reference to points of concentrated or discontinuous loading, by Louis Napoleon George Filon, 1903)

Predominantly, this kind of usage occurs in Series A of the Transactions "Containing Papers of a Mathematical or Physical Character", where it is highly formulaic.

Similarly to the semantic concept of ASSUMING in mathematics and related areas, MEASURING becomes an important methodological concept in many disciplines and we predominantly encounter *measuring* used as a gerund to form an adverbial of instrument—again a highly conventionalized usage (see example 3).

3) *It was found by […] measuring its distance from the nitrogen rays and from the two helium rays […]* (On the spectrum of the more volatile gases of atmospheric air, which are not condensed at the temperature of liquid hydrogen. – Preliminary notice, by George Downing Liveing and James Dewar, 1900)

leading appears conventionalized due to its use in *leading to*, both in concrete and abstract uses, often occurring after nouns. See examples 4 and 5.

[4] The entry actually quotes an attestation from the Philosophical Transactions: *She has a Navel-rupture, owing to the Ignorance of the Man in not applying a proper Bandage*. (Extracts of Two Letters from the Revd Dean Copping, F. R. S. to the President, concerning the Caesarian Operation Performed by an Ignorant Butcher; And concerning the Extraordinary Skeleton Mentioned in the Foregoing Article. By John Copping, 1739).

TABLE 5 | Three closest neighbours of VVGs with increasing Paradigmatic Variability (pvar+) by 50-year period.

Word	1675	1725	1775	1825	1875	1925
Adding	Add	Substituting	Inserting	Addition	Add	Introducing
	Subtracting	Add	Substituting	Applying	Dissolving	Dropping
	Substituting	Remembering	Subtracting	mixing	Introducing	Titrating
making	Make	Make	Make	Make	Make	Make
	Rendering	Performing	Performing	Obtaining	Rendering	Taking
	Completing	made	Pursuing	Bringing	Completing	Getting
Taking	Take	Take	Take	Take	Take	Take
	Took	Took	Took	Assuming	Took	Putting
	Putting	Selecting	Putting	making	Takes	making
Giving	Gives	Give	Give	Give	Gives	Gave
	Give	Gives	Gave	Gives	Give	Give
	Gave	Gave	Imparting	Gave	Gave	Gives
Obtaining	Attaining	Attaining	Attaining	Procuring	Getting	Securing
	Securing	Procuring	Determining	Discovering	Procuring	Getting
	Procuring	Deciding	Interpreting	Ascertaining	Preparing	Procuring

TABLE 6 | Three closest neighbours of VVG with stable Paradigmatic Variability (pvars) by 50-year period.

Word	1675	1725	1775	1825	1875	1925
According	Agreeably	Obeying	Conformably	Agreeably	Accordance	Accordance
	Conforming	Agreeably	Conformable	Conformably	Conformity	Ccording
	Conformable	Conformably	Agreeably	Conformity	Conformable	Irrespective
Regarding	Concerning	Concerning	Concerning	Respecting	Respecting	Concerning
	Attributing	Respecting	Deciding	Concerning	Concerning	Respecting
	Elucidating	Investigating	Estimating	Governing	Relating	Relating
Including	Excluding	Excluding	Comprising	Comprising	Comprising	Excluding
	Encircling	Replace	Excluding	Viz.	Excluding	Comprising
	Impressing	Forty-one	Besides	Excepting	Excepting	Excepting
Concerning	Regarding	Regarding	Respecting	Respecting	Respecting	Regarding
	Respecting	Respecting	Relating	Regarding	Regarding	Respecting
	Touching	Relating	‘On	Relating	Relating	Relating
Considering	Examining	Noticing	Contemplating	Reviewing	Discussing	Discussing
	Contemplating	Observing	Noticing	Consider	Consider	Consider
	Investigates	Experiencing	Re-examining	Conceive	Examining	Examining

4) *[. . .]Dr. Dunbar Hughes and Captain Calder started out along the road leading to the south* (Report on the eruptions of the soufrière, St. Vincent, 1902, and on a visit to Montagne Pelèe, in Martinique. -Part I. by Tempest Anderson and John Smith Flett, 1903)
5) *In discussing the results of the flash spectra obtained in India in 1898, I stated certain conclusions leading to the belief that the flash spectrum does, in fact, represent the upper more diffused portion of an absorbing stratum [. . .]* (Solar eclipse of 1900, May 28—General discussion of spectroscopic results, by John Evershed, 1903)

involving is predominantly used in postnominal position forming a reduced alternative to a relative clause (*which involves*). This usage thus appears highly conventionalized. See example 6.

6) *In the above deduction of such a law, we have used the general formulae involving sources of two types* (I. The integration of the equations of propagation of electric waves, by Augustus Edward Hough Love, 1901)

Similar patterns arise for the other pvar- verb forms, i.e., the past tense and past participle forms (VVD, VVN). The *ed*-form is a highly ambiguous form that is used for past tense, to form nonfinite adverbial clauses, as adjective as well as postmodifier (reduced relative clause). An example of an item that went a similar way as *owing (to)* is *provided*. Next to its lexical, verbal meaning, according to which it is used in past tense, active voice (example 7) or as postmodifier (example 8), it is used as a conjunction, in earlier usage with subjunctive mood (example 9). Our diachronic model clearly captures the shift towards the use of *provided* as a conjunction (as in 9) siding with other conjunctions such as *since* or *while* and landing in the same frequency range.

7) *I provided the best Opium I could get* (Of the Use of Opium among the Turks. By Dr. Edward Smyth, 1695)
8) *An assistant, provided with an apparatus, for writing down observations* (Description of a Forty-Feet Reflecting Telescope. By William Herschel, 1795)
9) *a most useful agent in separating olefiant gas from such mixtures, provided light be entirely excluded during its operation* (On the Aeriform Compounds of Charcoal and Hydrogen; With an Account of Some Additional Experiments on the Gases from Oil and from Coal. By William Henry, 1821)

4.3 Microanalysis: Items With Increasing Paradigmatic Variability

While the general diachronic trend is reduction, there is one set of items among the adverbs that actually expands. As mentioned in the introduction (**Section 1**), another characteristic trait of convergence over time is that items become conventionally associated with a particular meaning, e.g., occupying a predominantly interpersonal function (e.g., adverbs expressing stance) or a textual function (e.g., adverbs functioning as discourse connectors). In our data, this is the case for particular groups of adverbs which show increased variability (pvar+) coupled with decreasing mean distance and increased frequency over time. Adverbs within these groups become exchangeable, their neighbourhoods manifesting a continuous influx of new lexemes carrying similar interpersonal meaning, notably to express stance (*considerably*, *apparently*), or adopting similar textual functions, notably discourse markers (e.g., *thus*, *accordingly*).

Figure 9 shows six adverbs out of the top pvar+: three with textual and three with interpersonal meaning. Mean distance shows decreasing tendencies, i.e., neighbourhoods become semantically more coherent. As an example of textual meaning, **Table 7** shows decreasing mean distance and increasing variability for the neighbourhood of *thus*. While in the 18th century neighbours are semantically more varied with a mixture of textual and interpersonal meanings, by the 19th and 20th centuries, the textual meaning clearly prevails covering different kinds of semantic relation (e.g., concessive, temporal, adversative). Considering an example of interpersonal meaning, from **Table 8**, we see how *apparently* moves from a mixture of attitudinal (e.g. *dangerously*, *fatally*, *assuredly*) and epistemic meanings (e.g. *evidently*, *improbably*) to mainly the latter—a turn which seems to happen around the end of the 18th/beginning of the 19th century. We can observe that mean distance exhibits a rise by 1825 (from 0.32 to 0.43), where the epistemic *probably* is left as the only neighbour (at 0.6 distance threshold; cf. **Section 3.3**). In subsequent years, the neighbourhood around *apparently* is again further populated with other epistemic markers. Attitudinal markers are not included any more among the nearest neighbours and their mean distance to epistemic neighbours decreases (from 0.43 in 1825 to 0.33 in 1925). Thus, while paradigmatic variability increases, enriching the space with more items, mean distance to selected neighbours decreases. From a producer perspective, there is more choice but the meaning expressed is more specific (here: epistemic). Thus, expansion in types goes together with confinement in meaning, i.e. we encounter here conventionalization at the semantic level.

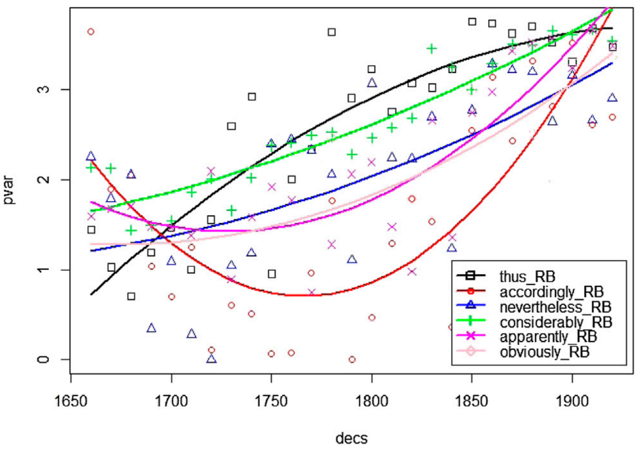

FIGURE 9 | Adverbs with increasing Paradigmatic Variability (pvar+).

5 SUMMARY AND CONCLUSION

We have explored the assumption that language use, while being under the permanent pressure of innovation, ultimately strives for conventionalization. The push for innovation is associated with cultural change and geared towards expressivity; the pull for conventionalization is language-internal and the optimization criterion is communicative utility. In our data, we observe for instance that the "chemical revolution" during the 18th and 19th centuries is linguistically reflected in temporary bursts of new terminology, e.g., associated with the oxygen theory of combustion that replaced the former phlogiston theory, indexed by a temporary rise in entropy for instance in the word cluster of terms for chemical elements. While innovation may thus result in temporary highs of linguistic variability, we have shown that as a general diachronic trend, variability is reduced resulting in fewer and/or more diversified linguistic options—see again the *size* and *molecule* examples in **Section 4.1** at the lexical level or the *ing*-forms as discussed in **Section 4.2** at the level of grammar.

Focusing on conventionalization, we have proposed a formal model of paradigmatic variability using word embeddings to represent the notion of paradigm by neighbourhood in vector space. The word embedding space is then analyzed in terms of diachronic change by systematically inspecting the (changing) neighbourhoods of words in terms of distance in vector space and entropy in a given neighbourhood. The overarching diachronic trend is a reduction of paradigmatic variability as shown by overall increasing distances between words and overall decreasing entropy. The observed entropy reduction is thus the measurable effect of a continuous, diachronic process that serves managing linguistic variability in the interest of rational communication. In the domain of discourse considered here—science—diversification in the lexico-semantic area is of course related to the evolution of scientific disciplines with their respective terminologies in the time period considered. Here, we do see temporary increases in paradigmatic variability (e.g., terms for chemical elements), but eventually it is pulled down again. In the lexico-grammatical area, we have seen that diversification is manifested by selected word forms leaving their lexical context, isolating themselves and/or landing in a grammatical usage context, i.e. they become function words (see the example of *owing (to)* in **Section 4.2**).

The only diachronic increase of paradigmatic variability was observed regarding specific adverbs with interpersonal meaning (stance, evaluation) or textual function (discourse connector) (as discussed in **Section 4.3**). This may lead to the interpretation that interpersonal and textual functions tend to give in more to the

TABLE 7 | Top neighbours (up to 30) for *thus* showing decreasing mean distance.

year	mdist	Neighbours
1675	0.40	So, mechanically, hereby, designedly
1725	0.41	So, demonstrably
1775	0.38	Now, then, mentally, eventually, hereby, previously, likewise, subsequently, sixthly, therefore, hereto, so, scrupulously, incidentally, prematurely
1825	0.30	Then, now, however, also, so, therefore, which, but, and, as, be, only, finally, yet, is, anyhow, intentionally, when, not, being, approximatively, statically, consequently, for, unnaturally, been, by
1875	0.35	Then, now, thereby, therefore, also, similarly, finally, again, hence, hereby, consequently, so, synthetically, accordingly, here, eventually, perforce
1925	0.33	Then, therefore, now, hence, consequently, thereby, also, finally, similarly, nevertheless, so, evidently, accordingly, sometimes, furthermore, subsequently, presumably, ultimately, again, which, indeed, likewise, eventually, i.e., but

TABLE 8 | Top neighbours (up tp 30) for *apparently* with increasing and decreasing mean distance.

year	mdist	Neighbours
1675	0.32	Unquestionably, evidently, whit, undoubtedly, essentially, improbably, scarcely, rarely, indubitable, dangerously, mostly, gravely, fatally, intrinsically, simply, oddly, seemingly, unobserved, drooping, questionable, assuredly, visibly, miraculous, doubtful, fundamentally, notoriously, preternaturally, soonest
1725	0.34	Demonstrably, essentially, invariably, improbably, unquestionably, unheard, inaccurately, remarkably, doubtfully, very, much, probably, confessedly, correctly, surely, indisputably, inconstancy, indubitably, incomparably, also, reality, gravely, obnoxious, only, immensely, conspicuously, hiss, receded, not
1775	0.40	Undoubtedly, obviously, probably, really, intrinsically, not, nominally
1825	0.43	Probably
1875	0.37	Probably, sometimes, possibly, evidently, essentially, presumably, undoubtedly, perhaps, almost, usually, physically, molecularly, doubtless, nearly, anyhow, occasionally, generally, likewise
1925	0.33	Probably, evidently, presumably, obviously, really, undoubtedly, doubtless, usually, certainly, possibly, practically, still, often, sometimes, generally, originally, necessarily, not, almost, also, invariably, always, actually, ordinarily

pressure of innovation/expressivity as a continuous trend, while the diachronic development in the ideational area exhibits only temporary rises in expressivity and a continuous pull towards conventionalization.[5] But this would warrant a dedicated empirical analysis in which interpersonal, textual and ideational functions are thoroughly separated. Yet another study would be warranted using data from “general language”, other domains or modes of discourse. First, to assess whether an item has grammaticalized or not, an important condition is that it spreads to other contexts. Second, scientific language is highly planned discourse between experts and will therefore exhibit fairly strong signals of communicative optimization. This may well be different in spoken contexts or in literary works. In fact, in a related study comparing the RSC with the Penn Parsed Corpus of Modern British English (PPCMBE), we found that only scientific texts show a significant diachronic trend towards dependency length minimization, which is considered another signal of communicative optimization (Juzek et al. (2020)). However, as we have shown, it is not only the reduction of options in context but also diversification of options that reduces entropy. In fact, diversification has been independently discussed as a general diachronic trend. For instance, an analysis of the 793,733 word forms included in the OED Historical Thesaurus reveals a strong diversification of vocabulary over the attested history of English, especially in the last two centuries, which is clearly due to the vast societal changes and technological advances in modern time.[6]

In terms of methods of diachronic analysis we presented a data-driven approach using as a basis a state-of-the-art computational language model. Apart from modeling words in their left and right context, the type of model employed—structured skip-gram word embeddings—enjoys the property of being aware of linear order. In this way, we not only pick up a lexical but also a grammatical signal. To evaluate diachronic changes we analyze the topology of the word embedding space as well as the entropy of words in their neighbourhoods. Entropy provides not only a diagnostic tool of diachronic change but gives us a direct link to a communicative interpretation of the observed diachronic patterns. Crucially, the proposed methodology allows us to track change by informational contribution rather than frequency alone. For instance, in our data it is primarily the mid-frequency items that are shown to be susceptible to change while high-frequency items are shown to be rather resilient. Many high-frequency words are already communicatively optimized—most function words have short codes and quite a few lexical words in the high frequency range are ambiguous/polyfunctional. Here, ambiguity can be considered

[5] Interestingly, related trends are observed in contact-induced language change by so-called borrowing hierarchies according to which textual (e.g., the connective *but*) and interpersonal items (e.g., modals) are most immediately affected (see e.g., Matras (2020)).

[6] see https://ht.ac.uk/treemaps/; Kay (2012).

another characteristic of code optimization, as shown by Piantadosi et al. (2012). While high(er) frequency of occurrence is thus not a condition of change, we can observe very clearly that frequency increase is a consequence of certain patterns of change, e.g. conventionalization by convergence/substitution (see again the *size* example in **Section 4.2**). Such observations are especially enabled by the methods and tools proposed here.

By high-level summary, we have shown that communicative concerns, as indexed by entropy, play an important role in the dynamics of language use, acting as a control on linguistic variability. The specific direction of research pursued here—the role of rational communication in linguistic variation and change—is in line with recent work on other aspects of language dynamics (e.g., language evolution Hahn et al. (2020)) and the specific approach proposed can be applied to other domains of inquiry where the interplay of communicative efficiency and socio-cultural change is involved, such as the linguistic dynamics in social media groups (e.g., Danescu-Niculescu-Mizil et al. (2013)) or the (changing) linguistic repertoires of individuals over a life time (e.g., Anthonissen and Petré (2019)).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Various word embedding models of the Royal Society Corpus with different parameter settings are made available from: http://corpora.ids-mannheim.de/openlab/diaviz1/description.html. A dedicated visualization of the models is made available publicly at: http://corpora.ids-mannheim.de/openlab/diaviz1/flying-bubbles.html#embeddings=rsc-diachron-1929-perplexity50-init-tc0-t1. The Royal Society Corpus 6.0 Open is available under a persistent identifier from: https://fedora.clarin-d.uni-saarland.de/rsc v6/.

AUTHOR CONTRIBUTIONS

ET developed the overall rationale of the study and carried out the analysis on reduced paradigmatic variability. PF trained the word embeddings and designed and implemented the diachronic analysis of paradigmatic variability. YB designed and implemented the analysis of diachronic semantic distances and helped with the collection of qualitative samples. SD-O carried out the analysis of items with increasing paradigmatic variability.

FUNDING

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 232722074–SFB 1102.

REFERENCES

Abnar, S., Ahmed, R., Mijnheer, M., and Zuidema, W. (2018). “Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity,” in Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018), Salt Lake City, UT, USA, January, 2018, 57–66.

Anthonissen, L., and Petré, P. (2019). Grammaticalization and the linguistic individual: new avenues in lifespan research. *Linguistics Vanguard.* 5, 20180037. doi:10.1515/lingvan-2018-0037

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., and Zeschel, A. (2010). Cognitive corpus linguistics: five points of debate on current theory and methodology. *Corpora.* 5, 1–27. doi:10.3366/cor.2010.0001

Asr, F. T., and Demberg, V. (2020). Interpretation of discourse connectives is probabilistic: evidence from the study of but and although. *Discourse Process.* 57, 376–399. doi:10.1080/0163853X.2019.1700760

Auguste, J., Rey, A., and Favre, B. (2017). “Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks,” in Proceedings of the 2nd workshop on evaluating vector space representations for NLP, Copenhagen, Denmark, September, 2017, 21–26.

Aylett, M., and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang. Speech.* 47, 31–56. doi:10.1177/00238309040470010201

Babanejad, N., Agrawal, A., An, A., and Papagelis, M. (2020). “A comprehensive analysis of preprocessing for word representation learning in affective tasks,” in Proceedings of the 58th annual meeting of the association for computational linguistics, July, 2020, 5799–5810.

Biber, D., and Gray, B. (2016). *Grammatical complexity in academic English: linguistic change in writing. Studies in English language*. Cambridge, UK: Cambridge University Press.

Bizzoni, Y., Degaetano-Ortlieb, S., Menzel, K., Krielke, P., and Teich, E. (2019). “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, Italy, August, 2019. Editors N. Tahmasebi, L. Borin, A. Jatowt, and Y. Xu (Stroudsburg, PA: Association for Computational Linguistics), 175–185.

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

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge: CUP.

Bybee, J., and Hopper, P. (2001). “Frequency and the Emergence of linguistic structure. No. 45,” in *Typological studies in language.* Editors J. L. Bybee and P. J. Hopper (Amsterdam/Philadelphia: John Benjamins).

Coles-Harris, E. H. (2017). Perspectives on the motivations for phonetic convergence. *Lang. Linguist. Compass.* 11 (12), e12268. doi:10.1111/lnc3.12268

Cornish, H., Dale, R., Kirby, K., and Christiansen, M. H. (2016). Sequence memory constraints give rise to language-like structure through iterated learning. *PLoS One* 12, e0168532. doi:10.1371/journal.pone.0168532

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 Brazil, Rio de Janeiro, Brazil, May, 2013. Editors D. Schwabe, V. Almeida, H. Glaser, R. Baeza-Yates, and S. Moon (New York, NY, US: Association for Computing Machinery). 3107–3318.

De Deyne, S., Perfors, A., and Navarro, D. (2018). “Learning word meaning with little means: an investigation into the inferential capacity of paradigmatic information,” in Proceedings of the 40th annual conference of the cognitive science society, Madison, WI, July, USA, 2018, 1608–1613.

De Smet, H. (2016). How gradual change progresses: the interaction between convention and innovation. *Lang. Var. Change* 28(1), 83–102. doi:10.1017/S0954394515000186

Degaetano-Ortlieb, S., and Piper, A. (2019). “The scientization of literary study,” in Proceedings of the 3rd joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature, Minneapolis, Minnesota, 7 June 2019. Editors B. Alex, S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz (USA: Association for Computational Linguistics), 18–28.

Degaetano-Ortlieb, S., and Teich, E. (2019). Toward an optimal code for communication: the case of scientific English. *Corpus Linguist. Linguistic Theory* [Epub ahead of print]. doi:10.1515/cllt-2018-0088

Delogu, F., Crocker, M., and Drenhaus, H. (2017). Teasing apart coercion and surprisal: evidence from ERPs and eye-movements. *Cognition* 161, 49–59. doi:10.1016/j.cognition.2016.12.017

Di Carlo, V., Bianchi, F., and Palmonari, M. (2019). "Training temporal word embeddings with a compass," in Proceedings of the AAAI conference on artificial intelligence, Honolulu, HI, USA, January, 2019, Vol. 33, 6326–6334.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2016). Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue Linguaggio*. 15, 7–28. doi:10.1418/83652

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, Denmark, September, 2017. Editors M. Palmer, R. Hwa, and S. Riedel (Copenhagen, Denmark: Association for Computational Linguistics), 1136–1145.

Eckart, R. (2012). "Grammaticalization and semantic re-analysis," in *Semantics. An international handbook of natural language meaning*. Editors C. Maienborn, K. von Heusinger, and P. Portner (Berlin: de Gruyter), 2675–2701.

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)*. July 2017. Editor N. Bubenhofer (Universität Zürich, Institut für Computerlinguistik, Zürcher Kompetenzzentrum Linguistik).

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 the 12th language resources and evaluation conference (LREC), May 2020. Editors N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. (Marseille, France: European Language Resources Association), 794–802.

Garrod, S., Tosi, A., and Pickering, M. J. (2018). "Alignment during interaction," in *Oxford handbook of psycholinguistics*. Editors S.-A. Rueschemeyer and M. G. Gaskell (Oxford, UK: OUP), Chap. 24.

Gessinger, I., Möbius, B., Andreeva, B., Raveh, E., and Steiner, I. (2019). "Phonetic accommodation in a wizard-of-oz experiment: intonation and segments," in Proceedings of interspeech 2019, Graz, Austria, September, 2019. Editors G. Kubin and Z. Kačič (Austria: Graz), 301–305.

Gries, S. T., and Hilpert, M. (2008). The identification of stages in diachronic data: variability-based Neighbor Clustering. *Corpora*. 3, 59–81. doi:10.3366/e1749503208000075

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 geometrical models for natural language semantics (GEMS 2011), EMNLP 2011, Edinburgh, United Kingdom, July, 2011. Editors S. Pado and Y. Peirsman (Stroudsburg, PA, US: ACL), 67–71.

Hahn, M., Jurafsky, D., and Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proc. Natl. Acad. Sci. U.S.A.* 117, 2347–2353. doi:10.1073/pnas.1910923117

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in Proceedings of the 2nd meeting of the north american chapter of the association for computational linguistics on language technologies (Stroudsburg, PA, US: ACL), 1–8.

Halliday, M. A. K. (1988). "On the language of physical science," in *Registers of written English: situational factors and linguistic features*. Editor M. Ghadessy (London: Pinter), 162–177.

Halliday, M., and Martin, J. (1993). *Writing science: literacy and discursive power*. London: Falmer Press.

Halliday, M. (1985). *Written and spoken language*. Melbourne: Deakin University 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 (EMNLP), November 2016. Editors S. Jian, D. Kevin, and C. Xavier (Austin, Texas: Association for Computational Linguistics), 2116–2121.

Harris, Z. (1991). *A theory of language and information. A mathematical approach*. Oxford: Clarendon Press.

Harris, Z. (2002). The structure of science information. *J. Biomed. Inf.* 35 (4), 215–221. doi:10.1016/S1532-0464(03)00011-X

Haspelmath, M. (1999). Why is grammaticalization irreversible?. *Linguistics* 37 (6), 1043–10680. doi:10.1515/ling.37.6.1043

Hawkins, R. D., Goodman, N. D., Goldberg, A. E., and Griffiths, T. L. (2020). Generalizing meanings from partners to populations: hierarchical inference supports convention formation on networks. arXiv:2002.01510.

Hilpert, M., and Perek, F. (2015). Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*. 1, 339–350. doi:10.1515/lingvan-2015-0013

Hollenstein, N., de la Torre, A., Langer, N., and Zhang, C. (2019). "CogniVal: a framework for cognitive word embedding evaluation," in Proceedings of the 23rd conference on computational natural language learning (CoNLL), Hongkong, China, November, 2019. Editors M. Bansal and A. Villavicencio (Hong Kong, China: Association for Computational Linguistics), 538–549.

Hume, E., and Mailhot, F. (2013). "The role of entropy and surprisal in phonologization and language change," in *Origins of sound change: approaches to phonologization*. Editor A. C. L. Yu (Oxford: Oxford University Press), 29–47.

Isbilen, E. S., and Christiansen, M. H. (2020). Chunk-based memory constraints on the cultural evolution of language. *Topics Cognit. Sci.* 12(2), 713–726. doi:10.1111/tops.12376

Jaeger, T. F., and Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. *Adv. Neural Inf. Process. Syst.* 19, 849–856. doi:10.7551/mitpress/7503.003.0111

Juzek, T. S., Krielke, P., and Teich, E. (2020). "Exploring diachronic syntactic shifts with dependency length: the case of scientific English," in Proceedings of universal dependencies workshop (UDW), coling, Barcelona, Spain, December, 2020, 109–119.

Kay, C. (2012). "The historical Thesaurus of the OED as a research tool," in *Current methods in historical semantics*. Editors K. Allan and J. Robinson (Berlin: Mouton de Gruyter), 41–58.

Kim, Y., Kim, K.-M., and Lee, S. (2020). "Adaptive compression of word embeddings," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, July, 2020, 3950–3959.

Kuperberg, G. R., and Jaeger, F. T. (2016). What do we mean by prediction in language comprehension?. *Cognit. Neurosci.* 31, 32–59. doi:10.1080/23273798.2015.1102299

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), Santa Fe, NM, August, 2018. Editors E. M. Bender, L. Derczynski, and P. Isabelle (Sante Fe, NM, USA:ACL), 1384–1397.

Labov, W. (1994). "Principles of linguistic change volume 1: internal factors," in *Language in society*. Editor P. Trudgill (Oxford: Blackwell Publishers).

Labov, W. (2001). "Principles of linguistic change volume 2: social factors," in *Language in society*. Editor P. Trudgill (Oxford: Blackwell Publishers).

Leech, G., Hundt, M., Mair, C., and Smith, N. (2009). *Change in contemporary English: a grammatical study*. Cambridge, UK: Cambridge University Press.

Lehmann, C. (1995). *Thoughts on grammaticalization*. München: Lincom.

Lemke, R., Horch, E., and Reich, I. (2017). "Optimal encoding!–information theory constrains article omission in newspaper headlines," in Proceedings of EACL 2017, April 2017. Editors L. Mirella, B. Phil, and K. Alexander (Valencia, Spain: Association for Computational Linguistics), 131–135.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian J. Linguist*. 20, 1–31.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi:10.1016/j.cognition.2007.05.006

Li, X. L., and Eisner, J. (2019). "Specializing word embeddings (for parsing) by information bottleneck," 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), Hongkong, China, November, 2019, 2744–2754.

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, June, 2015. Editors R. Mihalcea, J. Chai, and A. Sarkar (Denver, Colorado: Association for Computational Linguistics), 1299–1304.

Linzen, T., and Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cognit. Sci.* 40, 1382–1411. doi:10.1111/cogs.12274

Lowder, M. W., Choi, W., Ferreira, F., and Henderson, J. M. (2018). Lexical predictability during natural reading: effects of surprisal and entropy reduction. *Cognit. Sci.* 42, 1166–1183. doi:10.1111/cogs.12597

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition* 126 (3), 313–318. doi:10.1016/j.cognition.2012.09.010

Mair, C. (2017). "From priming to processing to frequency effects and grammaticalization? contracted semi-modals in present day English," in *The changing English language: psycholinguistic perspectives*. Editors M. Hundt, S. Mollin, and S. E. Pfenninger (Cambridge, UK: Cambridge University Press), 191–212.

Malisz, Z., Brand, E., Möbius, B., Oh, Y. M., and Andreeva, B. (2018). Dimensions of segmental variability: interaction of prosody and surprisal in six languages. *Front. Commun. Lang. Sci.* 3, 1–18. doi:10.3389/fcomm.2018.00025

Matras, Y. (2020). "Theorising language contact: from synchrony to diachrony," in *The handbook of historical linguistics of blackwell handbooks in linguistics*. Editors R. D. Janda, B. D. Joseph, and B. S. Vance (New Jersey: John Wiley & Sons Ltd), Vol. II, Chap. 18.

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*. Editor C. J. C. Burges, L. Bottou, and M. Welling (Red Hook, NY, USA: Curran Associates Inc.), 3111–3119.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* 320 (5880), 1191–1195. doi:10.1126/science.1152876

Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua* 108 (2–3), 95–117. doi:10.1016/S0024-3841(98)00046-1

Newmeyer, F. (2001). Deconstructing grammaticalization. *Lang. Sci.* 23 (2–3), 187–230. doi:10.1016/S0388-0001(00)00021-8

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: a case study. *Linguistics* 54 (1), 149–188. doi:10.1515/ling-2015-0043

Piantadosi, S., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition* 122 (3), 280–291. doi:10.1016/j.cognition.2011.10.004

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27 (2), 169–190. doi:10.1017/S0140525X04000056

Rothe, S., Ebert, S., and Schütze, H. (2016). "Ultradense word embeddings by orthogonal transformation," in Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies, San Diego, CA, USA, June, 2016, 767–777.

Schmid, H.-J. (2015). A blueprint of the entrenchment-and-conventionalization model. *Yearbook German Cognit. Linguist. Assoc* 3 (1), 3–26. doi:10.1515/gcla-2015-0002

Schulz, E., Oh, Y. M., Malisz, Z., Andreeva, B., and Möbius, B. (2016). "Impact of prosodic structure and information density on vowel space size," in *Proceedings of speech prosody*, Boston, June, 2016, 350–354.

Schwartz, D., and Mitchell, T. (2019). "Understanding language-elicited eeg data by predicting it from a fine-tuned language model," in Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Minneapolis, MN, USA, June, 2019, Vol. 1 , 2019 *(Long and Short Papers)*. 43–57.

Sikos, L., Greenberg, C., Drenhaus, H., and Crocker, M. (2017). "Information density of encodings: the role of syntactic variation in comprehension," in Proceedings of the 39th annual conference of the cognitive science society (CogSci 2017), November 2017 (London, UK: Curran Associates, Inc.), 3168–3173.

Speyer, A. (2007). *Germanische sprachen*. Göttingen: Vandenhoeck & Ruprecht.

Tourtouri, E., Delogu, F., Sikos, L., and Crocker, M. (2019). Rational over-specification in visually-situated comprehension and production. *J. Cultural Cognit. Sci.* 3, 175–202. doi:10.1007/s41809-019-00032-6

Traugott, E. C., and Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge: CUP.

Trudgill, P. (2008). Colonial dialect contact in the history of european languages: on the irrelevance of identity to new-dialect formation. *Lang. Soc.* 37 (02), 241–254. doi:10.1017/S0047404508080287

Ure, J. (1982). Introduction: approaches to the study of register range. *Int. J. Sociol. Lang.* 1982(35), 5–24. doi:10.1515/ijsl.1982.35.5

Venhuizen, N., Crocker, M. W., and Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy* 21 (12), 1159. doi:10.3390/e21121159

Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., and Wang, X. (2015). Word embedding composition for data imbalances in sentiment and emotion classification. *Cognit. Comput.* 7, 226–240. doi:10.1007/s12559-015-9319-y

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.